Gurram Lavanya, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

AI-Driven Drug Sensitivity Prediction Using COSMIC, DGIdb, and GDSC Integration

Gurram Lavanya[1], Ch.Naveen[2]

[1]Student, [2]Assistant Professor, Department of CSE,AI,ML, MAM Women's Engineering College Kesanupalli Narasaraopet

Abstract- It is critical to investigate the link between treatment efficacy and sensitivity to mutational patterns in order to successfully treat complex diseases such as cancer. Specifically, cancer drugs may lose some of their efficacy over time as a result of new mutations, and cancer cells themselves are continually changing as a result of ongoing mutations. The main purpose of this research is to analyse the relationship between medications, disorders, and genes using statistical methods. To further anticipate the drug sensitivity of cancer cells based on genomic changes, a generic processing pipeline and machine learning models were developed. This was accomplished by comparing an improved database to four well-known open-source databases: one that had information on drug sensitivity in cancer cell lines, one that contained resources for somatic mutation data, and the other that contained information on gene-drug interactions. Text encoding, filtering, and optimisation were among the preprocessing procedures used to provide a fresh, enhanced dataset for use in machine learning and statistical analysis. Statistics were run on the supplemented database to find out how much of an impact gene-drug interactions had on drug sensitivity. On the other hand, machine learning algorithms that have been trained on datasets of drug interactions or somatic mutations may be used to forecast drug sensitivity. Incorporating feature significance and ablation studies, the research aims to provide a thorough analysis of gene and pharmaceutical sensitivity. A constant R2 score of 0.73 across several data sources, on top of an R2 score of 0.91 in early testing, demonstrated good generalizability for the built pipeline. Ablation studies, statistical analysis, and machine learning all provide new perspectives to the field of drug sensitivity prediction.

Keywords: ML models, data manipulation, data integration, drug sensitivity, ML Z-score prediction.

I. INTRODUCTION

Recent thirty years have seen tremendous growth in computer technology, genomics, and molecular biology, all of which have provided the groundwork for groundbreaking advances in pharmaceutical research. Innovations in medication repurposing and discovery have been notably sped up by the merging of these domains. [1,2], and 3]. Drug repurposing and drug discovery have been the main areas of concentration in the field of drug studies as of late [4]. One of the main goals of medication repurposing is to find new therapeutic applications for existing pharmaceuticals, outside their original indications [4, 5]. By skipping over steps like discovery, preclinical testing, and production preparation, drug repurposing hopes to cut down on the time, money, and effort needed to generate a new medicine [6]. Medications include

Sildenafil, Thalidomide, and Minoxidil, which were formerly prescribed for one set of symptoms but are now used to treat another set of symptoms altogether [4, 7, 8]. The ideas of drug sensitivity and drug resistance will surface when we explore the subject of drug repurposing further. The term "drug resistance" describes the situation in which a medicine fails to have the desired therapeutic effect, often because cancer cells have developed ways to avoid or become immune to the treatment. Refer to references [9], [10]. Such conditions could make pharmacological treatments less effective. Due to the intricate interaction between the biological system and the medication molecules, the effectiveness and reaction of the medicine may vary across patients [11], [12].

Predicting the sensitivity effects of pharmaceuticals on patients may be achieved by collaboration between genetics and computer science research. This sensitivity information can then be used for drug repositioning. Research into drug response prediction is an emerging yet crucial area of study, especially for the creation of tailored methods of treating cancer [10, [12], [13], [14]. With the proliferation of pharmacogenomic data and the fast evolution of high-throughput technology, it may soon be much easier to choose pharmacological compounds for therapeutic application. New data sources have emerged as a result of recent discoveries on the interplay between genes, mutations, illnesses, and medications.

GDSC [15], COSMIC [16], and DGIdb [17] are examples of such repository. In addition, new avenues for study have opened up thanks to initiatives like the Cancer Genome Project (CGP), the Genomics of Drug and Cancer Therapeutics Response Portal (CTRP), and the Cancer Cell Line Encyclopaedia (CCLE) [18]. The pharmacological responses to cancer cell lines were examined by Zhang and colleagues [19] and Yang and colleagues [15] using the CCLE and CGP databases, respectively. Targeted gene sequencing panels may be developed using genomic data in oncology to detect tumor-specific genetic driver alterations. One of the most basic features of cancer is mutations, which interfere with the functioning of live cells [20], [21]. These alterations, known as driver mutations, account for the physiological processes by which cells develop into cancerous cancer cells [20]. At the same time, Wang and colleagues [22] combined pharmacogenomic information on substances' chemical characteristics with genomic changes associated with cancer. Their findings emphasise the need of methodical techniques that integrate several pharmacogenomic data sources to enhance the precision of therapy response prediction.

An algorithm for the prediction of drug-drug interactions (DDIs) was built using the DrugBank dataset. It was used to construct a recommender system. The model was updated with expert knowledge using specific regularisation procedures to improve the accuracy and reliability of interaction predictions [23], [24]. Another strategy for improving medication efficacy and safety is to examine drug-drug interactions (DDIs), which could compromise both. Using the DrugBank dataset, an ensemble stacking model was able to get 99% accuracy, indicating that deep learning may be used to make even greater improvements [25]. More and more pharmacogenomic studies are making use of GDSC and other open-source drug databases.

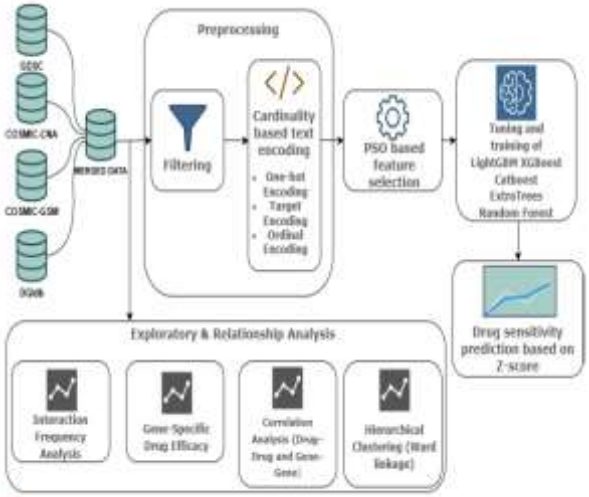
By combining gene expression, medication SMILES, and IC50 properties, for example, a new research used the GDSC dataset to build a glioblastoma drug effectiveness recommendation system. With RMSE = 0.98, MAE = 0.32, and R 2 = 0.90, the suggested convolutional neural network (CNN) outperformed the artificial neural network (ANN) model, which had RMSE = 1.21, MAE = 0.93, and R 2 = 0.87 [26].

Using cancer cell lines allows researchers to study how medications interact with genetic features including mutations and DNA methylation. A research by Iorio and colleagues [27] examined the genetic profiles of cell lines and how they responded to medications in a thorough manner. Researchers came to the conclusion that the genetic alterations seen in actual tumours, as replicated in cell lines, may influence the susceptibility or resistance of the cells to certain

medications. It is possible to utilise artificial intelligence to forecast pharmacological reaction without doing experiments. Long et al. [28] conducted an extensive analysis of the literature on artificial intelligence (AI) for the purpose of cancer medication resistance prediction. Carli et al. [14] used the GDSC and PRISM datasets to train a machine learning model that can understand how medications work and apply that knowledge to new patient samples. Learning from data is made possible by machine learning [29]. This area has been undergoing continuous development since the 1950s [30]. When given tabular data with several independent variables and a single dependent variable, machine learning algorithms may provide quite accurate results.

As a result, models for medication interactions that can learn from publicly available data are within reach. Using the GDSC dataset, Ha et al. [12] recently tested thirteen different regression models. Support vector regression (SVR) was determined to be the most effective method for producing the most accurate machine learning model. The significance of feature selection is emphasised by their findings. The correlation between medication resistance and mutational load was investigated in [31] using the GDSC dataset. To that end, they demonstrated how to create a drug resistance prediction model using Gradient Boosting. The research showed promise, but it might have been more robust and applicable if it hadn't relied on data from just one source.

Boosting in a functional space is the basis of Gradient Boosting, an alternative approach for machine learning. It builds a group of ineffective learners, usually basic decision trees [32]. To tackle classification and regression problems successfully, algorithms like LightGBM [33], XGBoost [34], and CatBoost [35] have been presented recently. Meng et al. [13] devised a method based on transfer learning to reduce the impact of dataset distributional discrepancies. They used massive datasets from places like GDSC and CCLE to train their model. In addition, Bueschbell et al. [36] investigated how AI and network biology contribute to our knowledge of cancer's multidrug resistance phenotype. The advent of AI and ML has revolutionised the way data is processed and examined.



medicine development [37]. Unfortunately, there is a lack of adequate evidence from which to draw conclusions about the complicated process behind cancer initiation, progression, and medication sensitivity. By examining data on medication resistance, gene-drug interactions, and alterations in genes, this research hopes to uncover new trends. In order to achieve this goal, relevant encoders have been included into the following databases: DGIdb for gene-drug interactions; COSMIC for mutation profiles and CNA data; and GDSC for drug sensitivity.

Appropriate data imputation and filtering methods are developed to tackle the challenges of merging four separate datasets. Using suitable encoding techniques, we digitalized the combined dataset. For feature selection, we employed Particle Swarm Optimisation (PSO). Then, in order to uncover the pharmacogenomic patterns at play, statistical studies were carried out, such as hierarchical clustering and drug-level association.

Drugs were grouped according to their associated gene response patterns using hierarchical clustering analysis and the enriched dataset. Furthermore, we assessed the efficiency of gene responses to various

medications. We trained several machine learning models—XGBoost, CatBoost, LightGBM, Random Forest, and Extra Trees—using the processed dataset. In light of this, we postulate that drug sensitivity prediction models derived from this integrated and optimization-driven strategy will be both strong and applicable to other contexts.

II. METHOD–INTEGRATION AND PREPROCESSING OF DATA SOURCES

Figure 1 shows the study's processing pathway

Figure 1 shows that one study's output investigates gene-drug interaction features using hierarchical clustering, grounded on Pearson correlation [38] and the Ward [39] approach; the other study's output concentrates on building a machine learning model to forecast drug z-score values. An initial step of the workflow is the incorporation of publicly available datasets from GDSC, COSMIC, and DGIdb. The combined dataset was then prepared for machine learning methods by applying filtering, text encoding, and metaheuristic feature selection. The last step was to train and fine-tune prediction models to estimate the Z-score, a measure of medication sensitivity.

Here, the Z-score is the standard deviation of a drug's IC50 value from the mean sensitivity; it's a way to quantify the drug's effectiveness across cancer cell lines. We used the following equation to determine it.

$$Z_i = \frac{X_i - \mu}{\sigma} \qquad (1)$$

for each set of cell lines, μ represents the average IC50 value, while σ stands for the associated standard deviation. If the Z-score is negative, then the medicine is more effective at inhibiting growth, and if it is positive, then the sensitivity is low. In this setting, a strong quantitative foundation for measuring medication potency, finding useful molecules, and directing possible therapeutic repurposing is provided by reliable Z-score prediction. The GDSC database provided the Z-score values used in this investigation, which were the basis for training and evaluating the models.

A. COLLECTION AND INTEGRATION OF OPEN-SOURCE DATASETS

The first stage of the research included merging four separate datasets from the GDSC, COSMIC, and DGIdb databases into a single, comprehensive file. The first step was to access the GDSC database for medication sensitivity information. In order to examine the consequences of somatic mutations and drug sensitivity concurrently, two distinct data sets were combined with the GDSC data, one of which was retrieved from the COSMIC database.

As a last step, the study drew on the DGIdb database to provide light on gene-drug interactions. What follows is a part outlining the specifics of our technique.

B. COLLECTING GDSC DATA TO EVALUATE DRUG SENSITIVITY

Resources of cell line anticancer medication sensitivity data are available in the GDSC database, which aims to uncover therapeutic biomarkers in cancer cells [15]. In the GDSC datasets you may find pharmacogenomic characteristics that show how various cancer cell lines react to different drugs. Prior to making any predictions about the drug sensitivity of cancer cell lines, the GDSC1 and GDSC2 datasets were acquired from the Genomics of Drug Sensitivity in Cancer (GDSC) project. Included in the combined dataset are important metrics including NLME_RESULT_ID, COSMIC_ID, CELL_LINE_NAME, DRUG_NAME, PUTATIVE_TARGET, PATHWAY_NAME, LN_IC50, AUC, RMSE, and Z_SCORE.

By bringing together data collected from different experimental settings, GDSC datasets increase data volume and guarantee variety. With this method of data management, a more stable and broadly applicable machine learning model may be trained.

C. INTEGRATION OF GDSC AND COSMIC CELL LINES PROJECT

The second step was to merge this new information with the Cell Lines Project_CompleteCNA_v101_GRCh38.tsv.gz file that was supplied by the COSMIC database, which is a catalogue of somatic mutations in cancer [16]. You

may remember that COSMIC is a database that provides full, vetted information on somatic mutations in human cancer.

The columns that make up this dataset are:
MINOR_ALLELE, MUT_TYPE, COSMIC_COV, COSMIC_SAMPLE_ID, SAMPLE_NAME, COSMIC_PHENOTYPE_ID, TOTAL_CN, and COSMIC_DELETE_CNV cols STUDY_ID, chromosome, genome, genome_start, genome_stop, and gene symbol. Copy number alterations, which might affect drug response mechanisms, can be evaluated at the gene level with the use of CNA characteristics. For instance, tumour suppressor gene deletions might lead to resistance patterns, but oncogene amplifications may be associated with enhanced treatment sensitivity.

The model's biological interpretability and robustness of downstream drug sensitivity predictions were enhanced by including these genome-wide CNA signals, which gave essential information on gene dosage effects. Two columns, COSMIC_GENE_ID in the CNA data and COSMIC_ID in the GDSC data, are designated as important characteristics in order to combine the two sets of information. The COSMIC Cell Line Project was also integrated to guarantee data variety and provide a more complete picture of the molecular features of cancer cell lines.

D. INTEGRATION OF DRUG-GENE INTERACTION DATABASE

Thirdly, the data was enhanced by using the interactions.tsv file from the DGIdb [17] database, which is a resource for druggable genomes and gene-related interactions. There are now additional columns in this dataset that include GENE_CLAIM_NAME, GENE_CONCEPT_ID, GENE_NAME, INTERACTION_SOURCE_DB_NAME, INTERACTION_SOURCE_DB_VERSION,
INTERACTION_TYPE, INTERACTION_SCORE, DRUG_CLAIM_NAME, DRUG_CONCEPT_ID, DRUG_NAME, APPROVED, IMMUNOTHERAPY, and ANTI_NEOPLASTIC. We compared the GENE_SYMBOL and DRUG_NAME columns in our dataset with the GENE_NAME and DRUG_NAME columns in the DGIdb dataset to complete the matching procedure. Through this procedure, we were able to identify drug-gene interactions of biological significance and, more specifically, possible interaction pairings related to cancer therapeutic targets. Nevertheless, owing to discrepancies in the names of some documents, it was necessary to eliminate them from further consideration.

E. INTEGRATION OF GENOME SCREEN MUTANTS FROM COSMIC DATABASE

The last step was to merge the COSMIC database's CellLinesProject_Genome
ScreensMutant_v101_GRCh38.tsv.gz file with the integrated dataset. The COSMIC_SAMPLE_ID and GENE_SYMBOL columns were used to match the somatic mutation information in this dataset with the current dataset. We used the following columns for downstream evaluation: COSMIC_SAMPLE_ID, GENE_SYMBOL, COSMIC_GENE_ID, TRANSCRIPT_ACCESSION, MUTATION_ID, MUTATION_DESCRIPTION, MUTATION_AA, MUTATION_CDS, LOH, HGVSP, HGVSC, GENOMIC_MUT_ALLELE, GENOMIC_WT_ALLELE, and MUTATION_SOMATIC_STATUS. This integration allows for a more thorough analysis of how individual mutations found in cell lines may impact medication sensitivity. The end product of these procedures was a complete and high-dimensional dataset that included genetic and pharmacological details. This dataset has laid a solid groundwork for the prediction of effective sensitivity measures like Z_SCORE and the discovery of possible pharmacological targets.

F. Data Preprocessing And Feature Selection

After integrating four datasets, cleansing the data to eliminate duplicates, mistakes, and unnecessary information is important to assure data integrity. Hence, it was important to remove irrelevant characteristics and duplicates from the dataset and focus on variables that really convey useful information. It was also necessary to deal with missing data. We were unable to find any pre-existing literature describing the necessary tools for combining the four files. So far, the processing pipeline has resulted in a modified dataset with 26 independent variables and 1 dependent variable.

The dataset includes characteristics from 4 distinct data sources.

Manipulation Of Categorical Data And Missing Fields

To make it easier to employ in statistical studies and machine learning models, numerical representations of the integrated data were considered beneficial. As a result, categorical characteristics were encoded using techniques determined by their cardinality. Depending on the number of classes (cardinality), several encoding techniques were used to digitise categorical data. Variables with low cardiacity were encoded using One-Hot, those with medium cardiacity with Target, and those with high cardiacity using Ordinal. For the same reason, we did not include columns with a large cardinality, such as genomic location or particular allele information, as they do not provide anything useful to numerical modelling.

Handling Missing Data

The existence of missing data is one of the downsides of integrating four separate databases. Statistical methods may be used to get around these obstacles. Making a well-rounded data structure that aids in model training and validation was, thus, the goal.

The integrated dataset underwent a thorough preparation and variable selection procedure prior to the application of machine learning. The dataset was cleansed first of duplicate columns with the _x and _y prefixes, as well as columns having redundant information due to data merging processes (e.g., GENE_NAME and GENE_CLAIM_NAME). Furthermore, metadata columns like COMPANY_ID, WEBRELEASE, and INSTITUTE were eliminated as they are not directly related to the model's predictive ability and are not important for analysis.

We isolated measures like RMSE, AUC, and LN_IC50 from the dataset because of their strong association with the objective variable, Z_SCORE. Avoiding degenerate learning behavior—in which the model would capture average drug response patterns rather than physiologically significant genomic interactions—made this exclusion all the more crucial. In order to prepare the data for training and validation of machine learning models, the explanatory variables were removed and the target variable (Z_SCORE) was isolated.

G. Exploratory Analysis Of Drug-Gene Interactions

Methods for examining drug-gene interactions and sensitivity to drugs are outlined here. The goal was to identify patterns that would indicate how genes and medicines interact in terms of sensitivity. Drug effectiveness evaluations made use of Z-scores. A significant inhibitory impact of the medicine on the target gene is shown by a negative Z-score, whereas a low effectiveness is indicated by a positive score [15].

Exploratory studies were conducted to investigate drug-gene interactions after the data integration step of the processing workflow. As part of the study, we count how many medications are linked to each gene and how many genes each drug targets. The number of encounters was used to measure the intensity of the interactions throughout the investigation.

Drug And Gene Clustering Analyses

This part of the research included grouping medicines and genes according to their z-score profiles. There was an assumption that medicines in the same cluster could have similar repurposing capabilities. When classifying medications based on gene expression and pharmacological reactions, Ward's minimal variance technique was favoured [39]. The following is the definition of Ward's minimal variance:

$$D(A,B) = \frac{|A||B|}{|A|+|B|}\|\bar{x}_A - \bar{x}_B\|^2 \qquad (2)$$

A and B are two clusters in Equation 2, where |A| and |B| are the sample numbers in each cluster, and xA and xB are the centroids of their respective clusters. Using this metric, we may combine the two clusters whose sum of within-cluster variation is increased by the least amount. Hierarchical clustering is used with gene and pharmacological activity profiles to improve similarity performance.

An strategy based on the Euclidean Distance for cluster merging was used to minimise total internal variation across clusters. The definition of the Euclidean Distance is:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (3)$$

Equation 3 uses the number of features (n) and the k-th feature value (xik) for points i and j, respectively. This formula measures the linear distance between two features in a space with n dimensions.

To determine the correct gene-drug interactions with regard to drug sensitivity, we calculated Pearson correlation coefficients [38]. A Pearson correlation is defined as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (4)$$

where Xi and Yi stand for paired observations, X and Y for their respective means; for instance, drug sensitivity values between genes and between medications. The direction and intensity of the linear correlation between the two variables are hinted at by the coefficient r, which may take values between -1 and +1. In order to find sets of genes or medications with comparable patterns, our approach grouped them according to their response profiles. The data are presented in a heatmap format to better demonstrate the abundance of gene-drug relationships. Evaluating the patterns of high-dimensional pharmacogenomic data was the systematic goal of these investigations.

Genes were also grouped according on how they responded to different medications. Using this method, we were able to classify patients into subsets with shared gene-level response patterns.

2) Enhancing Features Prior to Model Development

Particle Swarm Optimisation (PSO) has recently been shown to enhance machine learning performance by decreasing model bias and redundant features. In high-dimensional datasets, PSO aids in identifying the most relevant variables, which improves accuracy and avoids overfitting. For instance, research in the field of medicine and genomics has shown that using PSO-based feature selection may enhance the accuracy and stability of predictions [40], [41], [42].

To improve model performance and reduce bias in drug sensitivity prediction, this work used PSO as a feature reduction strategy. With our filtering method in place, the feature size was decreased to 27 by eliminating characteristics that do not help to learning. The Particle Swarm Optimisation (PSO) approach, which enhances the model's predictive potential, was used as a second filtering method to get the best features and evaluate the most informative subset from the filtered 27-feature data. In high-dimensional combinatorial optimisation problems, the PSO method is used to get solutions that are almost optimum on a global scale and to demonstrate robustness in the face of local optima [43]. According to the math, PSO revises the location and velocity of each particle in the following way:

$$v_i(t+1) = \omega v_i(t) + c_1 r_1(p_i - x_i(t)) + c_2 r_2(g - x_i(t)) \quad (5)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (6)$$

in which xi(t) and vi(t) represent the position and velocity of particle i at iteration t, ω controls the trade-off between exploration and exploitation, c1 and c2 are the coefficients for cognitive and social acceleration, and r1,r2 ∼ U(0, 1) are random numbers within the range of [0, 1]. Every particle revises its course according to its own optimal position pi and the optimal position on a global scale, denoted as g.

Forty particles and five hundred iterations of PSO were used in this investigation. Based on empirical data, the hyperparameters were tuned to:

The sentence states that the parameters ϋ = 0.9, c1 = 0.5, c2 = 0.3, k = 5, and p = 2, where k is the size of the neighbourhood and p is the Minkowski p-norm parameter used for distance calculation.

The steady optimisation behaviour across iterations was ensured by selecting parameter values that balanced convergence speed and population diversity. The Random Forest [44] regressor model was used to assess the performance of each candidate feature subset using the R 2 score under five-fold cross-validation. In the feature selection procedure based on PSO, the fit criteria was the R 2 metric.

The optimisation process identified and deleted characteristics from the dataset that were unnecessary or did not contribute significantly to the overall analysis. This left a more efficient and effective subset of 11 features. Table 1 summarises the most relevant characteristics for accurate Z-score prediction that were determined during the PSO-driven optimisation procedure. These characteristics improve the machine learning model's prediction accuracy by combining genomic, pharmacological, and interaction-level information.

H. Complete Set Of Data And Performance Measurements

In its original, unprocessed form, our dataset included 5,747 samples and 72 features; upon filtering, it reduced to 27 features. The z-score is created as the response feature, whereas 26 characteristics represent the input. If the R 2 score goes up, it means the model is becoming better at predicting the Z-score, which is a useful characteristic to have. Data on target genes, pathways, mutations, and other traits are jumbled together in these 27 features. Our method culminates in a stage when the 27 characteristics are reduced to the 11 most effective features.

Section I: The Evolution of Machine Learning Models Machine learning-based drug sensitivity prediction was the focus of the subsequent research phase. Various machine learning methods were used to create new models, which were then evaluated for accuracy.

Recognising Top-Notch Elements

The data set underwent feature selection in an effort to enhance model performance and forestall overfitting. To begin, we used the VarianceThreshold technique to extract the training data set of variables that had a constant or zero variance. This removed irrelevant variables from the model, such as those with a constant value across all instances.

After that, we used the f_regression test to find out which of the remaining variables were significantly related to the dependent variable (Z_SCORE).

TABLE 1. Columns of the PSO-selected feature subset. Most informative variables identified through Particle Swarm Optimization for accurate Z-score prediction.

Column name	Description	Type / Example value
CELL_LINE_NAME	Unique name of the cancer cell line used in pharmacogenomic screening	String (e.g., "A549")
SANGER_MODEL_ID	Identifier of the model assigned by the Sanger Institute	Integer (e.g., 1450)
TCGA_DESC	Cancer type or tissue origin based on TCGA classification	String (e.g., "LUAD")
DRUG_ID	Numerical identifier of the tested drug compound	Integer (e.g., 273)
PUTATIVE_TARGET	Primary molecular target of the drug	String (e.g., "EGFR")
PATHWAY_NAME	Biological pathway associated with the drug's mechanism of action	String (e.g., "PI3K/AKT Signaling")
TOTAL_CN	Total copy number variation count for the corresponding gene	Integer (e.g., 4)
MINOR_ALLELE	Count of the minor allele in copy number variation data	Integer (e.g., 1)
interaction_score	Confidence score of the drug–gene interaction derived from DGIdb	Float (e.g., 0.84)
interaction_type_missing	Encoded indicator showing whether the interaction type was missing	Binary (0 or 1)
LOH_1	Loss of heterozygosity indicator from genomic variant data	Categorical (e.g., "LOH present")

This test was used to determine the level of significance (p-value) and the variance contribution of each independent variable on the dependent variable. The variables that were included in the modelling procedure were those having a p-value less than 0.05, as determined by the findings. The model was trained using all available variables if no significant variables were discovered. Random Forest, Extra Trees, Lightgbm, catboost, and xgboost were among the methods used to forecast textual properties.

We trained five separate regression models utilising these eleven features: XGBoost[34], CatBoost[35], LightGBM[33], Random Forest[44], and Extra Trees [45]. Our selection of these tree-based models was based on their superior generalizability, robustness against missing data, and interpretability in complex and diverse biological datasets. As each algorithm ran, its most efficient settings were evaluated. Hyperparameter optimisations were carried out to achieve this goal. Using the RandomizedSearchCV approach, we optimised the hyperparameter values of each model [46]. The following parameters were evaluated in this context and optimised: XGBoost's n_estimators, max_depth, learning_rate, and colsample_bytree; CatBoost's

depth, learning_rate, and l2_leaf_reg; LightGBM's n_estimators, num_leaves, and learning_rate; Random Forest and Extra Trees' n_estimators, max_depth, and max_features.

As a result of this optimisation, the trained models were simpler, more accurate, and devoid of superfluous characteristics. J. Integrating CTRP-CCLE Data Integrating drug response data from the Cancer Therapeutics Response Portal (CTRP) [47] with multi-omics profiles from the Cancer Cell Line Encyclopaedia (CCLE) [48], the framework—originally developed for the GDSC setting—was further applied to an independent pharmacoge nomic context in order to perform a more comprehensive evaluation of the proposed pipeline.

The area under the dose-response curve (AUC) is the direct measure of drug response in CTRP, in contrast to GDSC, which often uses log-transformed IC50 values to depict drug sensitivity. DepMap identities were used to harmonise CTRP drug-cell line response profiles with CCLE genomic data. Without making any changes particular to the dataset, the data from CCLE on copy number variation and somatic mutations was combined with the data from CTRP responses. Then, the same feature engineering, encoding, and feature selection procedures were used, just as in the original pipeline that was based on GDSC.

III. FINDINGS FROM THE EXPERIMENT SUBJECT

A. Experimental Setup and Data Integration

Thanks to the integration, a multidimensional dataset with genetic and pharmacological characteristics was created. Python was the language of choice for all experiments. The model was developed using the scikit-learn, XGBoost, and CatBoost packages. There was no preprocessing or encoding done on the dataset before it was divided into 80% training and 20% testing. To ensure that no information was leaked, the training data was the only set that underwent any encoding operations, including target encoding.

Part B: Investigating Drug-Gene Interactions

1. Analysing Interaction Frequency The results showed that the genes that interacted with the greatest number of medications and the drugs that showed the most broad gene connections. The data show that of the drugs tested, the one with the most interactions is 273 (CUDC-101) with 1089, followed by 158 (PF-562271) with 520, and 1401 (AZD5438) with 359. The five medications that interact with the most genes are shown in Table 2. Looking at the gene level, the COSG115384 gene (TP53) has the highest number of medication interactions (868),

TABLE 2. Drugs that interact most with genes.

DRUG_ID	Number of Interactions	Drug Name
273	1089	CUDC-101
158	520	PF-562271
1401	359	AZD5438
1490	279	SN-38
1049	272	PD173074

TABLE 3. Genes that interact most with drugs.

COSMIC_GENE_ID	Number of Interactions	Gene Name
COSG115384	868	TP53
COSG73399	670	EGFR
COSG89395	495	ERBB2
COSG112841	357	HDAC9
COSG114519	348	FGFR1

EGFR (COSG73399) had 670 interactions, followed by ERBB2 (COSG89395) with 495. In Table 3, we can see which five genes have the most medication interactions. These results provide further evidence that drug-gene interactions may be important therapeutic targets in the fight against cancer.

2. Analysis of Drug Efficacy

According to the literature, a negative z-score indicates effective cell growth suppression, while a positive Z-score indicates ineffectiveness [15]. So, the Z-score may be used to determine how effective a medicine is. The top five genes with the greatest interaction frequency were subjected to a thorough pharmacological effectiveness investigation for this purpose. Figure 2 displays the results of determining gene-specific drug sensitivity

patterns and creating drug response profiles based on Z-score values for each gene. Analysis of the COSG115384 (TP53) gene revealed a wide range of pharmacological responses, as shown in Figure 2. Prior discussion has shown that the TP53 gene plays a pivotal role in cancer studies.

The results of the studies indicate that Drug 1490 (SN-38) is quite sensitive, with a Z-score of −1.5. The drug-sensitive profile with typically negative Z-score values and a Z-score of −1.0 against Drug 87 was shown by the COSG73399 (EGFR), the second most interacting gene in the studies. Likewise, a Z-score of -0.8 was used to introduce drug sensitivity between the COSG89395 (ERBB2) gene and Drug 87 (GW843682X). Figure 2 shows that the COSG112841 (HDAC9) gene had positive Z-score values against all of the medicines that were tested.

This data points to the gene's potential significance in the development of drug resistance. With a Z-score of −1.2, the COSG114519 (FGFR1) gene demonstrated a very sensitive reaction to Drug 1048 (Mirin).

According to Figure 2, there is a complicated and multifactorial structure to gene-drug interactions, and each gene shows a distinct pattern of response to various medications. It is possible that genes like COSG73399 (EGFR) and COSG89395 (ERBB2), which normally have negative Z-scores,

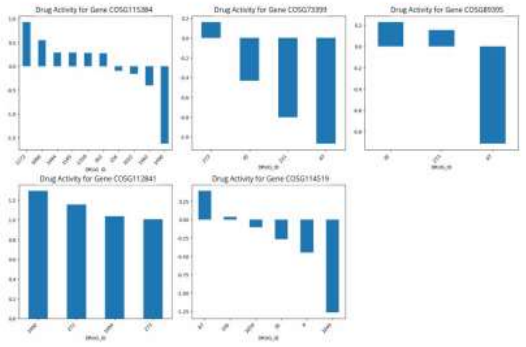


FIGURE 2. Gene-specific drug response profiles (Z-scores) for the 5 genes with the highest interaction frequency. Gene-specific drug response profiles (Z-scores) for the top five genes (TP53, EGFR, ERBB2, HDAC9, FGFR1). Negative Z-scores indicate higher drug efficacy, positive values lower sensitivity.

Efficient targets for cancer treatment. The COSG112841 (HDAC9) gene, on the other hand, has a positive Z-score across all medicines, which indicates that the gene in question could be crucial in drug resistance. Further evidence that the related medicine may have broad-spectrum anti-cancer action is the substantial negative impact of medicine 87 (GW843682X) on several genes.

C. Analysing Clusters and Correlations

The Outcomes of Hierarchical Clustering

The Ward technique was used to conduct hierarchical clustering analysis in order to categorise medications based on their gene response characteristics. Figures 3a and 3b show the resultant dendrogram, which used distance values to visually represent drug similarities and group medications based on their genetic response patterns. Figure 3a shows the pre-filtering clustering analysis, whereas Figure 3b shows the refined clusters after data preprocessing.

Figure 3b shows that out of the ten drug clusters identified by the revised clustering analysis, the most prominent of them was Clus ter 1, which included eight drugs: AT-7519, AZD5438, GSK690693, TAK-715, ZM447439, PI-103, PHA-793887, and OSI-930. These medications were found in close proximity to one another and had comparable impacts on gene response patterns, which might indicate that they have similar pharmacologic qualities or action mechanisms.

There were a total of ten clusters, each including a single medicine (PLX-4720, AZD8055, ZSTK474, SN-38, CI-1040, AZD6482, GW843682X, PF-562271, and NVP-TAE684), suggesting that these agents had different effectiveness profiles. To determine how genes react to various medications, researchers used hierarchical clustering analysis based on the Ward technique. Gene similarity distances are shown in the ensuing dendrograms (Figures 4a and 4b), which group profiles based on their drug response patterns.

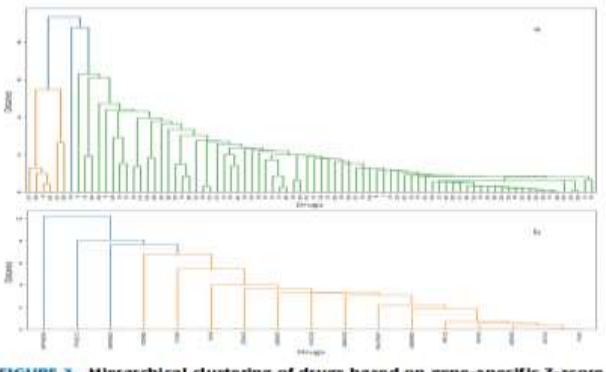


FIGURE 3. Hierarchical clustering of drugs based on gene-specific Z-score profiles using Ward's linkage and Euclidean distance, revealing clusters with similar pharmacogenomic responses. (a) Initial clustering analysis. (b) Refined clustering with quality control filtering (>5 genes per drug).

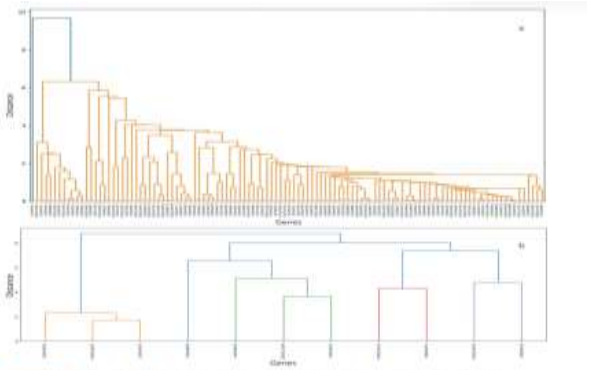


FIGURE 4. Hierarchical clustering of genes based on drug response Z-scores using Ward's linkage and Euclidean distance, showing functionally similar or co-regulated gene groups. (a) Initial clustering analysis. (b) Refined clustering with quality control filtering (>5 drugs per gene).

Both the original analysis (Figure 4a) and the updated technique with better quality control methods (Figure 4b) are shown. Figure 4b shows the results of the revised gene clustering study, which uncovered ten groups of genes. One cluster had two genes, COSG102697 and COSG91620, whereas the other clusters only included one gene. Genes with shorter distances on the distance axis react similarly throughout the medication panel, suggesting that their pharmacological effectiveness is comparable.

Based on these results, it seems that genes that are part of the same cluster have comparable resistance or sensitivity mechanisms or are involved in similar biological processes. 2) Results of the correlation analysis Analysis of drug-gene correlations allows for a thorough investigation of links based on effectiveness. Consequently, it gives crucial insights into the similarities and differences between the effectiveness responses of genes and medications, and it enables the identification of underlying biological interaction networks and pharmacological commonalities.

Colour coding based on correlation coefficients is used to display similarities in effectiveness profiles between medications in the drug correlation heat maps of Figure 5a and Figure 5b. Tones of red on the colour scale indicate positive correlations (similar effectiveness profiles), whereas tones of blue indicate negative correlations (opposite efficacy profiles). The level of resemblance in effectiveness between the variables is shown in this visualisation, where the correlation coefficients vary from -1 to +1.

Figure 5b shows the results of the updated study, which ranked the positive drug-drug correlations from greatest to weakest. After SN-38 and CUDC-101 (r = 0.9495, n = 3 shared genes), ZSTK474 and PI-103 (r = 0.8663, n = 3 shared genes), and TAK-715 and GW843682X (r = 0.9780, n = 3 shared genes), the following pairs of genes showed the strongest correlations. Furthermore, there were three other significant positive correlations: TAK-715 with NVP TAE684 (r = 0.8113, n = 4 shared genes), PHA-793887 with AZD5438 (r = 0.7975, n = 6 shared genes), and PHA-793887 with AT-7519 (r = 0.7969, n = 6 shared genes).

The existence of such robust positive correlations raises the possibility that different classes of drugs target similar biological processes and have comparable pharmacological effects. On the flip side, there were medications with contrasting patterns of effectiveness, as shown by the analysis's high negative correlations. The ZSTK474 and AZD6482 genes showed the largest negative connection (r = −0.9986, n = 3 shared genes), followed by the TAK-715 and PF-562271 genes (r = −0.6736, n = 5 shared genes), and finally, the SCH772984 and PLX-4720 genes (r = −0.3730, n = 3 shared genes). The medications with diverse pharmacodynamic properties or modes of action may have complimentary treatment methods or separate biological pathways, as the negative correlations imply contrary efficacy connections between them.

Figure 6a shows the results of the investigation that looked for correlations between genes. The study found that there were connections between the

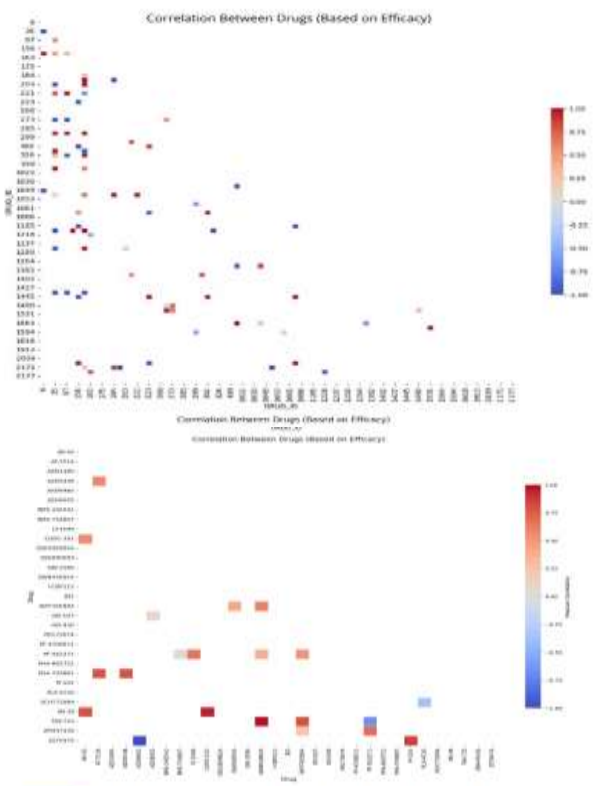


FIGURE 5. Drug efficacy correlation heat map. Red tones on the color scale represent positive correlations, while blue tones represent negative correlations. Dense red areas indicate consistency in drug efficacy profiles and similar pharmacological effects. (a) Initial correlation matrix. (b) Refined correlation matrix with quality control filtering (>3 shared genes) and statistical significance testing ($p < 0.05$).

We effectively discovered gene clusters with similar drug response patterns after evaluating the effectiveness responses of genes to various medicines. The top ten highest positive gene-gene correlations were discovered by the revised gene correlation analysis (Figure 6b). After COSG66634 and COSG58883 (r = 0.9721, n = 3 shared medications), COSG95525 and COSG82958 (r = 0.9567, n = 4 shared drugs), and COSG101532 and COSG58063 (r = 0.9857, n = 3 shared pharmaceuticals), the following pairs of drugs showed the strongest association. In addition, COSG95878 and COSG91620 had a significant positive correlation of 0.9324 (n = 3 shared pharmaceuticals) while COSG107313 and COSG106784 had a strong positive correlation of 0.9271 (n = 4 shared drugs).

Extremely highly correlated gene pairs may reflect functionally related genes with coordinated drug response patterns, share regulatory pathways, or be engaged in comparable biological processes. It was also shown that there were strong negative associations between genes. The pair of COSG96889 and COSG111160 had the largest negative association (r = −0.9992, n = 3 shared pharmaceuticals), followed by COSG91620 and COSG64043 (r = −0.9992, n = 3 shared drugs), and finally, COSG114519 and COSG107370 (r = −0.9984, n = 3 shared drugs). Additional significant negative correlations were seen between COSG66634 and COSG64752 (r = −0.9878, n = 3 shared medicines) and between COSG73035 and COSG101532 (r = −0.9812, n = 3 shared pharmaceuticals). Negatively correlated gene pairs may play separate or opposing roles in the mechanisms of drug response, since they may be involved in biological processes that are hostile to one another and produce opposite response patterns.

In particular, by choosing agents with complimentary processes (negative correlations) or by discovering treatment options for patients who acquire resistance (positive correlations), the found drug-drug correlations might influence rational drug combination design. The patterns of gene-gene correlations also provide the groundwork for precision oncology's multi-gene signature-based prediction models, which may divide patients into responder and non-responder categories and improve therapy choices and clinical outcomes.

D. Assessments of Machine Learning Models

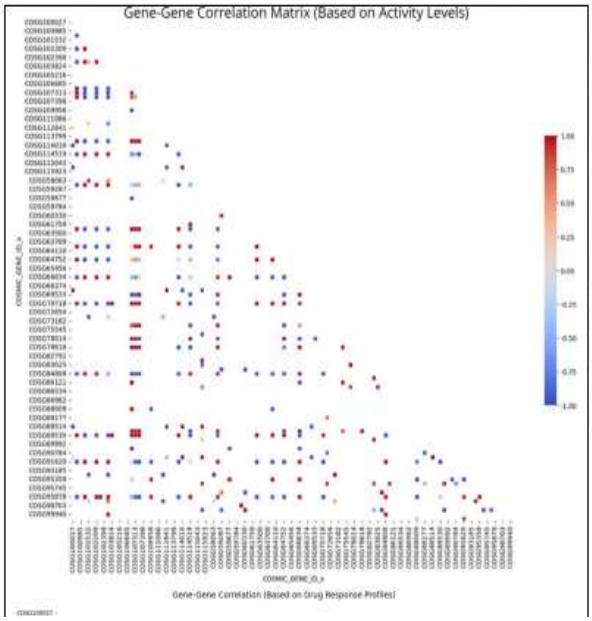
Here we evaluate and contrast the results of the machine learning models that we created. The following regression models are introduced in Table 4: LightGBM, XGBoost, CatBoost, Extra Trees, and Random Forest. In the table, we also provide our top hyperparameter results. The LightGBM Regressor model outperformed all others in terms of cross validation R 2 score, according to results from testing and training performance comparisons. So, to forecast Z_SCORE values for drug-gene interactions, Light GBM Regressor was chosen as the best model. Table 3 provides a summary of the optimal hyperparameters, together with the R 2 scores and other metrics, for each model.

Table 4 shows that out of all the models, LightGBM Regressor has the best R 2 (0.9120), making it the

most successful model. Simultaneously, the LightGBM Regressor offers the most modest MSE (0.1371) and MAE (0.1576). The XGBoost and CatBoost models were likewise very comparable to LightGBM in terms of performance, with R2 = 0.9110 and 0.9105 respectively, as well as MSE ~ 0.1385 and 0.1394, MAE ≈ 0.1816 and 0.1745, and AUC ≈ 0.9674 and 0.9671. Random Forest and Extra Trees were somewhat behind Random Forest and other error metrics with R 2 values, but their AUC values were in the 0.9666-0.9675 region, showing that they still had a lot of discriminative ability for the binary classification tasks. Despite the relatively great overall performance of

TABLE 4. Hyperparameter settings and cross-validation R^2 scores of trained models.

Model	Best Hyperparameters	R^2	MSE	MAE	AUC
LightGBM Regressor	learning_rate = 0.229; num_leaves = 40; n_estimators = 439	0.9120	0.1371	0.1576	0.9680
XGBoost Regressor	colsample_bytree = 0.77; learning_rate = 0.072; max_depth = 6; n_estimators = 408	0.9110	0.1385	0.1816	0.9674
CatBoost Regressor	depth = 6; l2_leaf_reg = 8; learning_rate = 0.189	0.9105	0.1394	0.1745	0.9671
Extra Trees Regressor	max_depth = 17; max_features = 0.421; n_estimators = 188	0.9082	0.1430	0.1642	0.9666
Random Forest Regressor	max_depth = 17; max_features = 0.421; n_estimators = 188	0.9066	0.1455	0.1755	0.9675



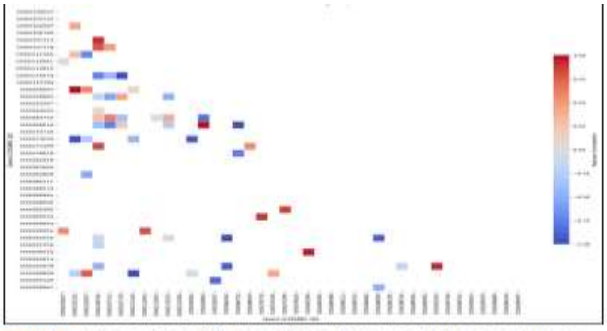


FIGURE 6. Intergenic activity correlation heat map. Red tones represent positive correlations, while blue tones represent negative correlations. Dense red areas indicate consistency in the activity response profiles of genes and the likelihood that they play a role in similar biological functions. (a) Initial correlation matrix. (b) Refined correlation matrix with quality control filtering (>3 shared drugs) and statistical significance testing ($p < 0.05$).

When it comes to consistently accurate predictions, LightGBM Regressor outperforms all tree-based models.

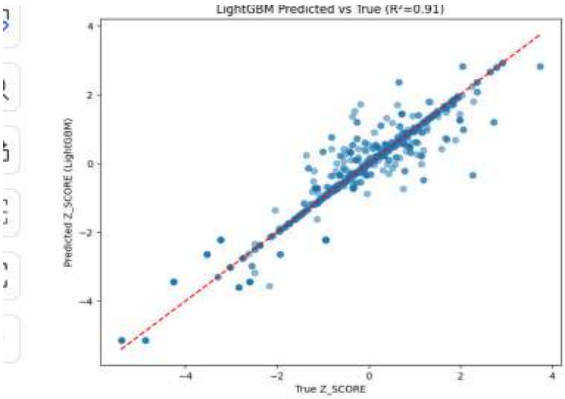


FIGURE 7. Predicted versus observed Z-scores for the LightGBM model (R^2=0.91). The close fit to the 1:1 line indicates strong predictive performance.

Further testing of the LightGBM Regressor is required due to its exceptional performance. The correlation between the observed and expected Z_SCORE values from the LightGBM model is shown in Figure 7. There is a strong clustering of the data points around the red dashed line, which stands for the ideal prediction line y = x; this indicates that the model has a high level of explanatory power (R 2 = 0.91). The consistent predictions from LightGBM for the most frequent observations in the dataset are shown by the dense distribution in the middle range, while the model's ability to generalise well even at edge values is shown by the small deviations at the extremities.

All things considered, the created LightGBM Regressor brings about peak performance. Figure 7

shows that LightGBM produces results that are near the best possible feasible prediction scores when taking the present dataset and its distribution into account. Five tree-based regression techniques (XGBoost, CatBoost, LightGBM, Random For est, and Extra Trees) are shown in Figure 8 together with their ROC curves, allowing for a comparison of the produced models for binary classification. Even with very low false positive rates, the model curves almost entirely overlap and attain true positive rates above 95%. Accumulated sensitivity (recall) and specificity are both shown by AUC values around 0.97. Simply put, the data contains a

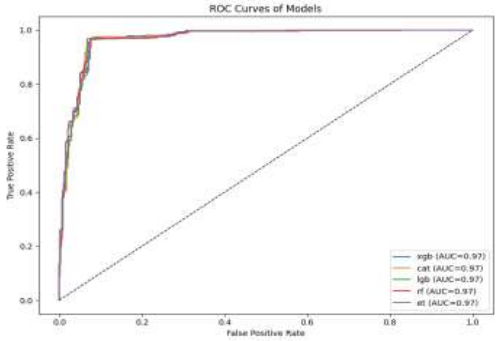


FIGURE 8. ROC curves of tree-based models (XGBoost, CatBoost, LightGBM, Random Forest, Extra Trees). All achieved high AUC values (~0.97), indicating strong discriminative performance.

structure that stands out in the high and low Z_SCORE categories to the degree feasible. The results show that the feature selection and preprocessing techniques used in Figure 8 provide strong generalisation capabilities in both continuous and binary tasks.

E. Ablation Study

We looked at four different ablation processes to see how characteristics affected learning capacity. That way, we can accurately gauge how the suggested preprocessing and feature selection procedures impacted the final model's performance. You can see the overall impact on the study's methodology and results from each part of the processing pipeline in Figure 9. Models from XGBoost, CatBoost, LightGBM, Random Forest, and Extra Trees were trained using default hyperparameters in each process, and then tested on an independent test set using R2, MSE, MAE, and AUC metrics. The models were optimised using RandomizedSearchCV.

Prior to VarianceThreshold, f_regression, and fixed PSO feature selection, the first pipeline employs one-hot encoding for variables with low cardinality, target encoding for variables with medium cardinality, and ordinal encoding for variables with high cardinality. While 11 features derived from PSO were used after VarianceThreshold in the PSO only procedure, the Freg only strategy just utilised VarianceThreshold and f_regression tests. Without applying any further feature selection phase, the Simple Encode method transformed all category variables just using label encoding. Figure 9 displays the R2 ratings of the ablation pathways.

At R 2 = 0.9120, MSE = 0.1371 and MAE = 0.1576 and AUC = 0.9680, the optimised LightGBM model outperformed the rest of the pipeline in the original setup. R 2 values of 0.9110 and 0.9105 were achieved by the optimised XGBoost and CatBoost models, respectively. R2 = 0.8824, MSE = 0.1833, MAE = 0.2366, and AUC = 0.9614 show that the optimised XGBoost model's performance was poorer in the Freg-only scenario compared to the original pipeline.

Optimisations for XGBoost (R2 = 0.9110), Extra Trees (R2 = 0.9101), and LightGBM (R2 = 0.9081) were seen in the PSO-only pipeline, which produced results that were comparable to the Original pipeline. As the most fundamental transformation approach, Simple Encode outperformed all other models with respect to XGBoost R2 = 0.8746, MSE = 0.1953, MAE = 0.2557, and AUC = 0.9602. For bioinformatics interaction data to be accurate and generalizable, our results show that multi-stage feature selection and extensive encoding approaches are essential.

F. Review of Related Research

By comparing it to other previous research that used the GDSC dataset, we were able to further validate the performance of the proposed model. While most prior research has used GDSC data (as seen in Table 6), our work intends to use GDSC, COSMIC, and DGIdb to increase feature variety and enhance the biological context. The table provides an explanation of the predictor factors utilised in each research and indicates which studies have the

ability to predict either LN_IC50 or Z-score. The R 2 column was used to assess the models' results.

The suggested LightGBM-based model outperformed when taking R2 performances into account; Table 5 shows that it attained the maximum accuracy with R2 = 0.91. The findings show that drug sensitivity predictions are improved when several source pharmacogenomic datasets are integrated. In addition, the findings point to the possibility of an efficient method based on PSO-based feature selection. Section G: Assessment Using Separate CTRP-CCLE Pharmacogenomic Data We utilised the comprehensive modelling framework to evaluate the proposed pipeline's generalizability beyond commonly used pharmacogenomic benchmarks.

This involved combining multi-omics data from the Cancer Cell Line Encyclopaedia (CCLE) with drug response profiles from the Cancer Therapeutics Response Portal (CTRP). Unlike datasets like GDSC, which often use log-transformed IC50 values to depict drug sensitivity, the CTRP dataset presents drug response as the area under the dose-response curve (AUC). In order to keep the original CTRP response scale, the prediction task was built on drug sensitivity, which is defined as SENS = -AUC. Without making any adjustments based on the dataset, the whole pipeline was used, which includes building genomic features, using leakage-free category encoding techniques, selecting features using metaheuristics, and carrying out gradient boosting regression. Both were used to summarise the genomic information from CCLE.

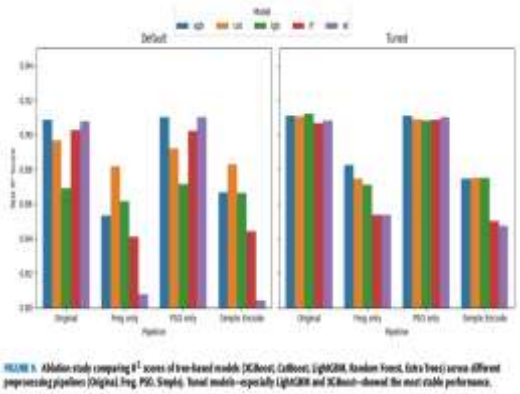


TABLE 5. Model performance on CTRP–CCLE integration using AUC-based drug sensitivity.

Target	R^2	MSE	MAE	Feature Set	Final Features
AUC	0.7376	1.7672	0.9229	43	13

aggregate descriptors and wide-summary statistics, with regulated encodings acquired from the training data solely using medication and tissue information. The LightGBM model got an impressive R2 score of 0.74 by using a subset of characteristics that were informative and chosen by Particle Swarm Optimisation (PSO) from a larger pool of numerical data. When merging the CTRP and CCLE datasets, our processing pipeline algorithm picked 13 characteristics at random. These features include drug identifiers (DRUG_NAME, TCGA_DESC), several Copy Number Variation (CNV) metrics (cnv_amp_minus_del, cnv_burden, cnv_mean_wide, cnv_q90, cnv_q90_q10_range, log1p_cnv_del_lt_0_90), mutational profiles (mut_sum_wide, mut_nonzero_genes), and statistical descriptors of the transcriptomic landscape (tcga_mean, dt_mean, dt_n). Here is a summary of the model that was generated: Table 5.

With an R2 value of around 0.74, the suggested model can account for a large chunk of the variation in absolute sensitivity to drugs across different cancer cell lines and chemicals. Using a very compact final feature set, this performance was achieved, which is a positive sign that the PSO-based feature selection technique may limit model complexity while helping to uncover meaningful genomic and contextual data.

TABLE 6. Comparison of recent studies using the GDSC dataset for drug sensitivity prediction.

Study	Dataset	Algorithm	Pred. Var.	R^2
Chawla *et al.* (2022) [50]	GDSC / CCLE	Deep Learning	LN_IC50	0.77
Schlüter and Schönhuth (2025) [51]	GDSC / CMP	Ridge Regression	IC50	0.87
Li *et al.* (2024) [52]	GDSC / CCLE	Deep Learning	LN_IC50	0.88
Yu and Fan *et al.* (2025) [53]	GDSC	Deep Learning	LN_IC50	0.75
Proposed model	**GDSC / COSMIC / DGIdb**	**LightGBM (Machine Learning)**	**LN_IC50 / Z-Score**	**0.91**

IV. DISCUSSIONS

In order to build a machine learning model to correctly predict medicine sensitivity, a large and comprehensive dataset is necessary. Integrating many data sources is a sensible and practical way to enhance databases.
Consequently, four independent datasets were collected and amalgamated: DGIdb, COSMIC's Copy Number Analysis dataset, GDSC, and the Cancer Drug Sensitivity database.

The many types of information offered by each dataset allow for comprehensive data analysis.

There are IC50, AUC, and Z-score metrics in the Drug Sensitivity dataset for a number of cancer cell lines that have been treated with known drugs. Annotations of somatic mutations are available in Genome Screen Mutants, whereas information on copy number alterations in cancer genomes is provided by Copy Number Analysis (CNA).

The last step was to use curated gene-drug interaction data from the DGIdb database to associate pharmacogenomic pro files with recognised biological targets. There are a number of problems that develop with enhanced dataset integration when the size of the combined dataset becomes more big and sparse. Additionally, several columns were either missing or duplicate data. This necessitates a multi-stage filtering procedure to address missing data and remove noise. In order to eliminate unnecessary features, we deleted columns with too much missing data and imputed others using appropriate statistical methods. Redundant identifiers, such as dataset IDs, institution names, and release information, were removed, and attributes with low variance were removed to further improve data consistency and reduce the risk of overfitting.

To eliminate multicollinearity, we used correlation-based filtering while retaining features with strong and independent information. The next step was to conduct feature importance analysis to identify the true discriminating variables. The filtering approaches inevitably reduced the number of usable characteristics from 72 to 27. The most valuable characteristics were then extracted from the revised dataset using a Particle Swarm Optimisation (PSO)–based feature selection method. This optimisation method automatically reduced the number of characteristics from 27 to 11, allowing machine learning to discover the optimal subset for Z-score prediction. After biomedical text data was changed using proper encoding and PSO-driven feature optimisation, the model's prediction performance was significantly enhanced, according to the PSO experiments. The research made a significant contribution to the field by introducing a new processing pipeline that improved data collection, combination, filtering, and the building of machine learning models. Several studies have made use of the GDSC, COSMIC, and DGIdb databases; nevertheless, this pipeline stands out due to its methodical processing, which entails merging several sources into one dataset, subsequently optimising features with great

precision. Then, we used the combined and enhanced data in our machine learning model building process. The improved dataset was able to learn medication sensitivity from the right databases, which allowed it to make accurate predictions. Using prospective clinical research to validate the models is essential before evaluating their effectiveness in a real-world clinical setting. One way to improve the models' generalizability is to test them on other cancer kinds and patient populations. From a translational medicine perspective, translating data from lab-grown cancer cell lines into clinically applicable prediction models is an important first step in bridging the gap between research and patient care.

Our study offers a new perspective on translational medicine in response to this interest. Results from analyses of gene-drug interaction patterns demonstrated a consistent relationship between TP53, EGFR, and ERBB2 and medication efficacy across a variety of compounds. These findings suggest that these genes may hold promise as potential therapeutic targets or markers for treatment response prediction. Clustering and

correlation analysis also revealed that drugs with the same target pathways had similar response signatures, which might lead to more rational combination therapies. With the right combination of computational insights and experimental confirmation, finding therapeutically relevant targets and bolstering tailored therapy regimens for cancer should be simpler. There is a growing interest in pharmacogenomics as a discipline that seeks to understand the genetic influences on pharmacological effects. Our method has the potential to make it easier to put this information into effect in the medical field. If we could predict which drugs would have a better chance of success in cancer types with certain mutation patterns, it would greatly aid in reducing the time and money needed for drug development. Using data collected from cancer cell lines in the lab to create prediction models that can be used in the clinic is a key objective of translational medicine.

The study has several limitations, despite the excellent findings. The present open-source data sources still do not provide accurate drug sensitivity predictions. Estimating the link between drug sensitivity and cancer sensitivity is becoming an increasingly popular topic of research.

Additionally, different databases have different amounts of data about cancer tissues. However, combining the sources for the sake of enrichment is no easy feat. Here, COSMIC details cancer tissue somatic mutations, and GDSC data illustrates how drug sensitivity differs among cancer cell lines. But DGIdb gives you the lowdown on gene-drug interactions. The computational difficulty may be further complicated by experimental data from newly created drugs. We need to put in more time and effort to ensure that the machine learning models we construct are robust against overfitting and can generalise effectively.

If the proposed model produces an abnormally high R2 value, it is important to take the experimental design into account while trying to understand it. All tests were performed in a controlled environment using cross-validation to ensure that no data leaked from the training set into the testing set. Because of the inherent complexity and noise in pharmacogenomic data, this result does not reflect an absolute measure of drug sensitivity predicting; rather, it represents the effectiveness of the recommended modelling approach within this situation. These findings show that optimised machine learning approaches may successfully anticipate patterns of pharmacological responses in controlled laboratory conditions. To ensure that the proposed methodology is applicable in other contexts, we transferred the pipeline from the combined GDSC-COSMIC-DGIdb architecture to a separate phar macogenomic scenario that made use of the CTRP and CCLE datasets. Rather than using GDSC's normalised measurements generated from LN_IC50, which normally characterise drug sensitivity, we adjusted the modelling procedure to accommodate the CTRP dataset's original label definition of drug response, which is area under t

he dose-response curve (AUC). To illustrate the mutation profiles and complete copy number variation derived from CCLE, distributional encodings and genomic aggregate characteristics were used. Integrating medication and tissue data with regulated encoding methodologies learnt from training data alone helped reduce information leakage. Next, the final prediction model was trained using LightGBM. A subset of numerical characteristics, chosen by Particle Swarm Optimisation (PSO) from a wider specified pool, was then considered informative. Results obtained with the CTRP-CCLE integration imply that the proposed approach may be effectively adjusted to accommodate various pharmacogenomic datasets with varying feature designs and drug sensitivity requirements. In order to overcome the limitations of our study and expand upon our conclusions, future research might use larger and more diverse data sets to enhance the generalizability of our models.

This will broaden the range of patients that our models may be used to in clinical practice. Genomic data must be integrated with transcriptomic, proteomic, metabolomic, and epigenomic data in order to comprehend the processes that influence

drug response. In future studies, we want to build more complex models using similar approaches. Use of explainable AI (XAI) methods has the potential to make machine learning models easier to understand and work with. Techniques such as SHAP (SHap ley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) may be used to explain future research predictions for clinical practice. Gene, mutation, and drug interaction databases are updated in real-time upon discovery of new experimental findings. Additionally, new working groups are making accessible additional data sources. Our approach will need frequent upgrades to include new information since open-source databases are dynamic. The incorporation of multi-omics data and other databases is one possible future strategy. An additional potential outcome of this endeavour may be the development of interactive tools and user-friendly interfaces. Making it easier for computational pharmacogenomic analysis and experimental data to communicate is the end objective of this study.

V. CONCLUSION

The researchers in this work used filtering and data categorisation procedures on open-source drug-gene-pathway-mutation datasets to create a new dataset on drug sensitivity and resistance. Statistical analysis was carried out after data enrichment in order to determine which medicines and genes were most linked with drug sensitivity. The results showed that there are strong correlations between medication sensitivity and gene mutation-drug interactions. Data collection and filtering followed a set of predetermined protocols throughout the study. Several feature cardinality algorithms are used to the combined dataset to guarantee fresh, qualified data.

Furthermore, we demonstrate how optimisation methods may be used to derive the most efficient features for data reduction. Using integrated pharmacogenomic data from GDSC, COSMIC, and DGIdb, this work proved that machine learning is an excellent tool for drug sensitivity prediction. After a number of models were trained and fine-tuned using hyperparameters, LightGBM proved to be the most effective in terms of predicted accuracy (R2 value). To further comprehend the connections between genes and medications, we not only built models but also performed hierarchical clustering, correlation analysis, interaction frequency assessment, and drug effectiveness profiling. Important trends emerged from these analyses, which could lend credence to investigations into precision medicine and medication repurposing in the future. Developing personalised medicine requires a thorough understanding of how gene mutations affect medication sensitivity. There is a current lack of AI-generated progress in the sector since data sources are inadequate in comparison to the possible alterations. The use of AI in personalised medicine is anticipated to provide better outcomes as the quantity of data continues to grow.

REFERENCES

1. B. Shaker, K. M. Tran, C. Jung, and D. Na, ''Introduction of advanced methods for structure-based drug discovery,'' Current Bioinf., vol. 16, no. 3, pp. 351–363, Mar. 2021.
2. B. Shaker, S. Ahmad, J. Lee, C. Jung, and D. Na, ''In silico methods and tools for drug discovery,'' Comput. Biol. Med., vol. 137, Oct. 2021,Art. no. 104851.
3. X. Zeng, F. Wang, Y. Luo, S.-G. Kang, J. Tang, F. C. Lightstone, E. F. Fang, W. Cornell, R. Nussinov, and F. Cheng, ''Deep generative molecular design reshapes drug discovery,'' Cell Rep. Med., vol. 3, no. 12, Dec. 2022, Art. no. 100794.
4. Z. Wan, X. Sun, Y. Li, T. Chu, X. Hao, Y. Cao, and P. Zhang, ''Applications of artificial intelligence in drug repurposing,'' Adv. Sci., vol. 12, no. 14, 2025, Art. no. e2411325.
5. S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, and M. Pirmohamed, ''Drug repurposing: Progress, challenges and recommendations,'' Nature Rev. Drug Discovery, vol. 18, no. 1, pp. 41–58, Jan. 2019.

6. T. T. Ashburn and K. B. Thor, ''Drug repositioning: Identifying and developing new uses for existing drugs,'' Nature Rev. Drug Discovery, vol. 3, no. 8, pp. 673–683, Aug. 2004.
7. I. Cg, J. Languillon, K. Ramanujam, G. Tarabini-Castellani, J. T. De las Aguas, B. Lm, K. Uemura, M. S. C. Domínguez, and T. Sundaresan, ''WHO co-ordinated short-term double-blind trial with thalidomide in the treatment of acute lepra reactions in male lepromatous patients,'' Bull. World Health Org., vol. 45, no. 6, pp. 719–32, 1971.
8. Convit, S. G. Browne, J. Languillon, J. H. Pettit, K. Ramanujam, F. Sagher, J. Sheskin, G. Tarabini-Castellani, L. de Souza Lima, J. G. Tolentino, M. F. Waters, L. M. Bechelli, and V. M. Domínguez, ''Therapy of leprosy,'' Bull. World Health Org., vol. 42, no. 5, pp. 667–72, 1970.
9. N. Vasan, J. Baselga, and D. M. Hyman, ''A view on drug resistance in cancer,'' Nature, vol. 575, no. 7782, pp. 299–309, Nov. 2019.
10. Y. Mao, D. Shangguan, Q. Huang, L. Xiao, D. Cao, H. Zhou, and Y.-K. Wang, ''Emerging artificial intelligence-driven precision therapies in tumor drug resistance: Recent advances, opportunities, and challenges,'' Mol. Cancer, vol. 24, no. 1, p. 123, Apr. 2025.
11. L. Wang, H. L. McLeod, and R. M. Weinshilboum, ''Genomics and drug response,'' New England J. Med., vol. 364, no. 12, pp. 1144–1153, 2011.
12. S. Ha, J. Park, and K. Jo, ''Comparative analysis of regression algorithms for drug response prediction using GDSC dataset,'' BMC Res. Notes, vol. 18, no. 1, p. 10, Jan. 2025.
13. [13] W. Meng, X. Xu, Z. Xiao, L. Gao, and L. Yu, ''Cancer drug sensitivity prediction based on deep transfer learning,'' Int. J. Mol. Sci., vol. 26, no. 6, p. 2468, Mar. 2025.
14. [14] F. Carli et al., ''Learning and actioning general principles of cancer cell drug sensitivity,'' Nature Commun., vol. 16, no. 1, p. 1654, Feb. 2025.
15. W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, ''Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells,'' Nucleic Acids Res., vol. 41, no. D1, pp. D955–D961, Nov. 2012.
16. S. A. Forbes, D. Beare, K. Gunasekaran, G. Keller, E. J. Lameijer, R. Luo, G. Turner, N. Bindal, B. Buckton, E. J. Coker, I. Cree, M. Dunlop, S. Enright, E. Finn, M. Jang, K. Lawrence, S. Mclaren, M. Oakley, C. Perez-Llamas, A. Prole, Y. Tang, G. Varese, C. Wan, H. M. Wood, C. Yong, F. Zhung, M. R. Stratton, and P. J. Campbell, ''Cosmic: Mining cancer genome data worldwide,'' Nucleic Acids Res., vol. 47, no. D1, pp. D787–D793, 2019.
17. M. Cannon, J. Stevenson, K. Stahl, R. Basu, A. Coffman, S. Kiwala, J. F. McMichael, K. Kuzma, D. Morrissey, K. Cotto, E. R. Mardis, L. Griffith, M. Griffith, and A. H. Wagner, ''DGIdb 5.0: Rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms,'' Nucleic Acids Res., vol. 52, no. D1, pp. D1227–D1235, Jan. 2024, doi: 10.1093/nar/gkad1040.
18. M. J. Garnett et al., ''Systematic identification of genomic markers of drug sensitivity in cancer cells,'' Nature, vol. 483, no. 7391, pp. 570–575, 2012.
19. N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, ''Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model,'' PLOS Comput. Biol., vol. 11, no. 9, Sep. 2015, Art. no. e1004498.
20. H. Najgebauer, U. Perron, and F. Iorio, ''Redefining false discoveries in cancer data analyses,'' Nature Comput. Sci., vol. 1, no. 1, pp. 22–23, Jan. 2021.
21. M. R. Stratton, P. J. Campbell, and P. A. Futreal, ''The cancer genome,'' Nature, vol. 458, no. 7239, pp. 719–724, 2009.
22. Y. Wang, J. Fang, and S. Chen, ''Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties,'' Sci. Rep., vol. 6, no. 1, p. 32679, Sep. 2016.
23. S. Jain, E. Chouzenoux, K. Kumar, and A. Majumdar, ''Graph regularized probabilistic matrix factorization for drug-drug interactions

prediction,'' IEEE J. Biomed. Health Informat., vol. 27, no. 5, pp. 2565–2574, May 2023.
24. G. Kumar, S. Yadav, A. Mukherjee, V. Hassija, and M. Guizani, ''Recent advances in quantum computing for drug discovery and development,'' IEEE Access, vol. 12, pp. 64491–64509, 2024.
25. S. Abbas, G. A. Sampedro, M. Abisado, A. S. Almadhor, T.-H. Kim, and M. M. Zaidi, ''A novel drug-drug indicator dataset and ensemble stacking model for detection and classification of drug-drug interaction indicators,'' IEEE Access, vol. 11, pp. 101525–101536, 2023.