2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822128

Predicting Drug Resistance in Cancer Cell Lines Using Machine Learning Approaches

Helena Hoy Loon Tam, Yu Mei Ng, Muhammad Azhar Ali Boolaky, Huey Fang Ong*
School of Information Technology
Monash University Malaysia
Bandar Sunway, Selangor, Malaysia
ong.hueyfang@monash.edu

***Abstract*—Cancers are the second leading cause of death worldwide, with breast cancer being the most prevalent type, accounting for 2.26 million cases in 2020. A major factor contributing to high cancer mortality rates is the ability of cancer cells to develop resistance to conventional therapies, highlighting the urgent need for further research and the development of more effective treatments. This prevalence led us to focus on predicting drug resistance in cancer cell lines, particularly breast cancer, using machine learning approaches. We propose a curated, comprehensive, and integrated dataset comprising gene expression profiles from the Cancer Cell Line Encyclopedia (CCLE), drug response data from the Genomics of Drug Sensitivity in Cancer (GDSC), and detailed drug information from PubChem. A notable advancement in our approach is the utilisation of the Morgan Fingerprint technique to effectively encode chemical compound structures to improve model compatibility and prediction outcomes. Furthermore, we employed the Deep Neural Network (DNN), which has demonstrated substantial improvements over existing models by achieving a root mean square error (RMSE) below 1. The results of our experiments underscore the model's robust predictive capability and adaptability to new data.**

Keywords—drug resistance, cancer cell line, machine learning, prediction, deep neural network

I. Introduction

The global burden of cancer has significantly increased, with nearly 10 million deaths in 2020 [1]. The Human Genome Project provided a comprehensive map of the human genome, revolutionising oncology by identifying genes linked to chemoresistance in cancer cells [2]. Despite the advancements in cancer treatments, the relationship between cancer cell molecular profiles and drug responses remains complex [3]. The persistent challenge of drug resistance in certain cancer cell lines remains a significant concern, rendering treatment failure.

Therefore, identifying and addressing drug resistance is crucial for effective treatments. Databases like the Genomics of Drug Sensitivity in Cancer (GDSC), Cancer Cell Line Encyclopedia (CCLE), and Simplified Molecular Input Line Entry System (SMILES) offer insights into cell line drug sensitivity and gene expression. Leveraging these datasets, researchers can investigate the relationship between gene expression and drug resistance in cancer cell lines and develop predictive models for drug resistance prediction.

Machine learning emerges as a promising approach to address the complexity of predicting drug resistance, enabling the creation of models capable of identifying cancer cell lines prone to developing resistance to specific drugs. By employing regression, classification, deep learning, and comprehensive data analysis techniques, these models offer valuable insights into drug resistance patterns and relationships, thereby enhancing the effectiveness and personalisation of cancer treatment [4]. Consequently, classification and calculation of the probability of drug resistance across various aspects can be conducted, enhancing the effectiveness and personalisation of cancer treatment [5].

However, substantial opportunities for improvement in terms of predictive performance and model generalisability still exist within these computational models [6, 7]. Traditional machine learning techniques face several challenges, including the curse of dimensionality [8], class imbalance [9], heterogeneity [10, 11], and the need for robust and interpretable outcomes [12]. These challenges hinder accurate drug resistance prediction, causing delays in effective cancer treatment. However, these limitations also present an opportunity to enhance prediction models by integrating diverse data sources and applying advanced machine learning techniques.

Hence, this paper aims to harness machine learning approaches to bolster our ability to predict and combat drug resistance in cancer cell lines, thereby enhancing the efficacy of cancer treatment and patient care. The resulting models are anticipated to serve as valuable tools for medical professionals and researchers, facilitating the development of more effective cancer therapies and treatments while minimising the risk of treatment failure or delays in disease management. The remaining content of this paper is organised as follows: Section 2 discusses some of the related works; Section 3 introduces the proposed method; Section 4 presents the experimental results; and finally, the conclusion is given in Section 5.

II. Related Work

Drug resistance, initially observed in antibiotic-resistant bacteria, is the domain factor affecting effectiveness in cancer treatment. While many cancers initially respond to chemotherapy, they can develop resistance due to factors such as DNA mutations and metabolic changes that promote drug inhibition and degradation [13]. Recent research has identified additional mechanisms behind drug resistance in tumours,

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

including tumour heterogeneity, cellular-level changes, and genetic factors [14]. The Human Genome Project, a ground-breaking effort, revolutionised cancer research by generating vast amounts of genomics data, mainly through technologies like next-generation sequencing [15]. These technologies facilitate comprehensive molecular profiling of cancer cell lines, particularly gene expression analysis. This wealth of genomic information is vital for identifying genes involved in cancer cell chemoresistance processes [16].

The success of the Human Genome Project has also provided access to critical datasets, such as the GDSC database [17], which has significantly advanced research on drug resistance. By analysing the gene expressions of cancer cells and their responses to various drugs, researchers can identify patterns of drug resistance. The GDSC database offers valuable information on cell-line drug resistance, genomic datasets, and genomic features related to drug sensitivity. Additional resources, such as the CCLE, SMILES, and PubChem, are also crucial for predicting drug resistance. These datasets, derived from high-throughput screenings by institutions like the Wellcome Trust Sanger Institute, cover over 1,000 cell lines [18].

With these datasets, researchers can explore the relationships between gene expression and drug resistance in cancer cell lines, enabling the development of predictive models for drug resistance. Various computational methods for predicting drug responses and identifying biomarkers have been extensively discussed. Machine learning techniques have emerged as powerful tools for these predictions, including Support Vector Machines [19], Bayesian multitask multiple kernel learning [20, 21], Random Forests [22-24], and neural network models [25]. By preprocessing data on cancer cells and drug information to serve as the features to train the predictive model via different variations of computational approaches, these approaches can determine drug response, predicting whether a drug will be sensitive or resistant to a particular cancer cell. The rationale for this study stems from identified limitations in existing models:

- Curse of Dimensionality Challenges [8]: Traditional machine learning techniques struggle to handle high-dimensional data efficiently, leading to increased computational complexity and potential overfitting. To address this, data-driven feature selection techniques, such as heuristic searches, regularisation [26, 27] or autoencoder [28] selection within learning algorithms, can be employed [29]. Deep learning presents a promising approach to managing high-dimensional data efficiently.
- Class Imbalance [14] and Data Heterogeneity [16]: Class imbalance occurs when the distribution of classes (e.g., drug-sensitive vs drug-resistant) in the dataset is highly skewed, leading to biased model performance. Data heterogeneity refers to variations and inconsistencies in the characteristics of the data, which can undermine the generalisability and effectiveness of machine learning models. Leveraging a pan-genome framework [30], which integrates heterogeneous and large-scale datasets, can enhance predictive capabilities. Additionally, incorporating different data types, such as gene expression and drug information specific to particular cancer types, can help identify heterogeneity among cancer classes and forecast drug resistance more accurately [31]. Focusing on specific cancer types, such as breast cancer initially instead of categorising all cancer types as one for training, allows for increased accuracy in drug prediction within the same cancer type, with potential expansion to other cancer types in the future, overcoming data heterogeneity.
- Lack of Robustness and Interpretability [12, 32]: Combining complementary machine learning models, as demonstrated by [33], has been shown to bolster the predictive prowess of drug response prediction models and enhance their robustness [20]. This effect is particularly pronounced when employing deep learning methodologies compared to conventional machine learning approaches, as highlighted by [7]. Additionally, existing models often lack interpretability in their outcomes, and there is a scarcity of efficient tools for predicting drug resistance in cancer cell lines. Hence, there is an opportunity to develop a model capable of generating interpretable outcomes by indicating whether the drug is resistant to the cancer cells through its prediction output.

III. Methodology

The development of the proposed drug resistance prediction model includes several main phases: data collection, data preprocessing, predictive model training and testing. The overall design of our training dataset is illustrated in Fig. 1. It comprises 6,847 rows and 19,482 columns. It includes three main features: 19,225 columns for gene expressions, 256 columns for 256-bit Morgan Fingerprint, and one target feature, the LN_IC50.

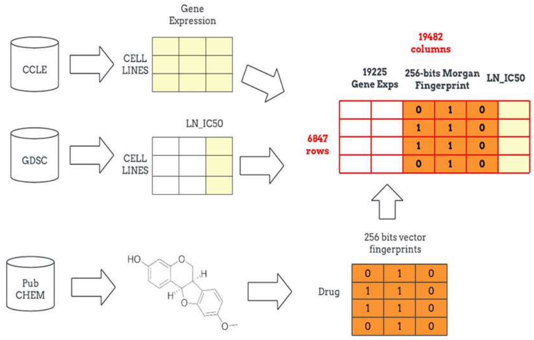


Fig. 1. Dataset Design and Preprocessing

A. CCLE Dataset Preprocessing

In the first phase, we focused on preprocessing the CCLE gene expression dataset. Key preprocessing steps included extracting essential features such as cosmic ID, DepMap ID, CCLE name, primary disease, and gene expression. These features served specific purposes: gene expressions formed the training data for model training, cosmic ID and DepMap ID were common features for merging with other datasets, CCLE name identified each cosmic ID, and the primary disease feature enabled the extraction of breast cancer-specific data to meet our study requirements. Columns were renamed for consistency with other datasets, facilitating easier merging with the GDSC dataset. Unnecessary columns were removed to prevent overfitting and improve model performance. Data cleaning was critical to remove rows with missing values, ensuring consistency and accuracy in the dataset. The phase concluded with the extraction of data specific to breast cancer.

B. *GDSC Dataset Preprocessing*

In the second phase, we processed the GDSC drug response dataset. This involved extracting features such as drug ID, cosmic ID, drug name, and LN_IC50. LN_IC50 was designated as the target feature for model training, while drug ID and cosmic ID were used for merging datasets, and drug name facilitated easy identification of each drug. Similar to phase one, unnecessary columns were removed, and data cleaning was performed to eliminate inconsistencies. The output of phase one was merged with the output of phase two on cosmic ID, resulting in the "Gene-Drug" dataset.

C. *PubCHEM Dataset Preprocessing*

The third phase involved preprocessing the PubCHEM dataset. Drug IDs from the "Gene-Drug" dataset were exported to the PubCHEM website to collect relevant data on drug compound information, specifically the isosmiles. Data belonging to GDSC1 were excluded, focusing on the GDSC2 dataset. Key preprocessing steps included extracting drug ID, PubCHEM ID, and drug Isosmiles. Isosmiles data were used for model training by converting them into binary representations via Morgan Fingerprint application, and PubCHEM ID and drug ID facilitated merging with other datasets. Columns were renamed, and data were cleaned, following the same rationale as in previous phases. Duplicate data were removed to ensure dataset consistency and reliability. Finally, the "Gene-Drug" dataset was merged with the output of phase three, resulting in the final "Gene-Drug-Chem" dataset, which was ready for model training.

D. *Predictive Model Building*

We have explored other machine learning approaches, such as Support Vector Regression (SVR), and settled on the Deep Neural Network (DNN) to design the model training and testing as shown in Fig. 2. The preprocessed "Gene-Drug-ChemStructure" dataset is used to build and evaluate a supervised predictive model. It is divided into 90% training data and 10% testing data using the 'train_test_split' function from the scikit-learn library. Further, the training data is split into 90% training data and 10% validation data. This ensures the model's ability to generalise and perform accurately with new, unseen data, which is crucial for real-world applications where users upload novel data for prediction. The objective is to verify if the model can make valid predictions for new cell lines or drugs not included in the training set, ensuring that the model is learning patterns and not merely memorising the training data, thereby enhancing its applicability in diverse, real-world scenarios. Before training, the features of the data are normalised using 'StandardScaler()', which standardises the features by removing the mean of each column and scaling them to unit variance.

As shown in the figure, we began the predictive model development by building a baseline DNN using the Sequential model from Keras. Initially, this model featured two layers plus an output layer. The first layer had 64 neurons, and the second had 128 neurons, both utilising the ReLU activation function. We chose the 'Adam' optimiser for its efficient handling of large datasets and adaptive learning rate properties and 'mean_squared_error' as the loss function due to its effectiveness in regression tasks by minimising the squared differences between predicted and actual values.

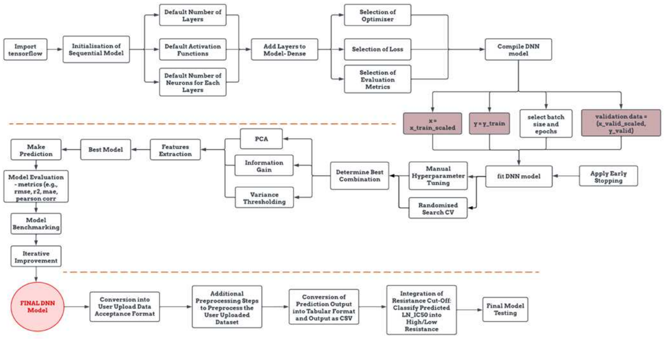
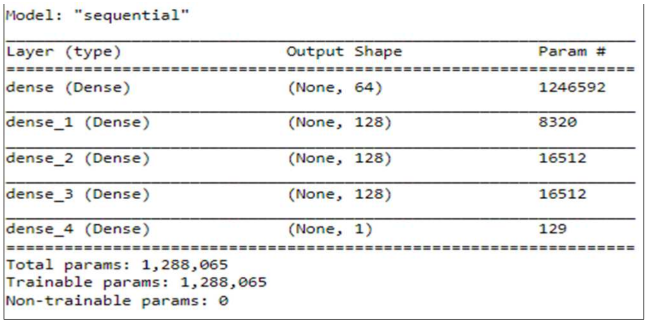

Import tensorflow
Initialisation of Sequential Model
Default Number of Layers
Default Activation Functions
Default Number of Neurons for Each Layers
Add Layers to Model- Dense
Selection of Optimizer
Selection of Loss
Selection of Evaluation Metrics
Compile DNN model
x = x_train_scaled
y = y_train
select batch size and epochs
validation data = (x_valid_scaled, y_valid)
fit DNN model
Apply Early Stopping
Manual Hyperparameter Tuning
Randomized Search CV
Determine Best Combination
PCA
Information Gain
Variance Thresholding
Features Extraction
Best Model
Make Prediction
Model Evaluation - metrics (e.g., rmse, r2, mae, pearson corr
Model Benchmarking
Iterative Improvement
FINAL DNN Model
Conversion into User Upload Data Acceptance Format
Additional Preprocessing Steps to Preprocess the User Uploaded Dataset
Conversion of Prediction Output into Tabular Format and Output as CSV
Integration of Resistance Cut-Off: Classify Predicted LN_IC50 into High/Low Resistance
Final Model Testing

Fig. 2. Design for Predictive Model Development

After compiling the baseline model, we trained it using the training and validation data. We also implemented early stopping to prevent overfitting and to restore the best weights based on the validation loss. Early stopping was crucial as it allowed us to halt training once the model's performance ceased to improve, thus saving computational resources and avoiding overfitting.

Observing the initial results, we proceeded with hyperparameter tuning, experimenting with various configurations of neurons and activation functions across different layers and adjusting the number of epochs and batch sizes. This comprehensive search aimed to identify the optimal combination of parameters. Additionally, we employed several feature selection techniques—such as Principal Component Analysis (PCA), Information Gain, Randomised Search CV, and Variance Thresholding, to reduce the dataset's high dimensionality. This step was necessary to manage the extensive gene expression data effectively. Using the refined model with selected features, we trained it on the testing data to predict the LN_IC50 values. We evaluated the model's performance with metrics like RMSE, R2, MAE, and Pearson correlation. Based on these evaluations, we iteratively improved the model, ultimately identifying the best DNN configuration that achieved the lowest RMSE.

With the final model, we converted it to a format compatible with user-uploaded datasets for integration into the web application. The user-uploaded dataset serves as the testing dataset, allowing predictions to be made. This necessitated additional preprocessing steps in the model code, as the user-uploaded dataset is not scaled, lacks Morgan fingerprints, and contains unnecessary columns. After preprocessing the user-uploaded dataset, the model can seamlessly make predictions on the data. To ensure the prediction results are interpretable within the web application, the prediction output is presented in a tabular format. This table includes the predicted LN_IC50 values for each drug-cell pair. To enhance user understanding, a resistance cut-off point is applied to the prediction results, classifying the drugs into categories of high or low resistance based on their LN_IC50 values.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                Output Shape              Param #
=================================================================
dense (Dense)               (None, 64)                1246592
_________________________________________________________________
dense_1 (Dense)             (None, 128)               8320
_________________________________________________________________
dense_2 (Dense)             (None, 128)               16512
_________________________________________________________________
dense_3 (Dense)             (None, 128)               16512
_________________________________________________________________
dense_4 (Dense)             (None, 1)                 129
=================================================================
Total params: 1,288,065
Trainable params: 1,288,065
Non-trainable params: 0
```

Fig. 3. The Architecture of the Proposed DNN Model

The architecture of our final DNN model, as depicted in Fig. 3., consists of a sequential arrangement of layers. The model features four dense (fully connected) layers followed by one output layer. The first dense layer has 64 neurons, while the second to the fourth layer has 128 neurons. All these dense layers use the ReLU activation function to introduce non-linearity into the model. The final output layer has a single neuron. This configuration allows the model to learn complex patterns in the data, making it suitable for predicting drug sensitivity.

IV. Experiments

A. Data sources

1) Genomics of Drug Sensitivity in Cancer (GDSC): The GDSC database, particularly the GDSC2-dataset, offers extensive drug response data for 969 cancer cell lines and 295 drugs. It is sourced from the Wellcome Sanger Institute and is among the largest cancer drug screening collections available [34]. Other than drug data featuring cancer cell line names, drug names, and drug responses indicating resistance, it also measures basal gene expression levels using microarray technology, enabling integration with gene expression datasets to predict drug sensitivity in normal and tumour tissues [35, 36] High-throughput drug screening technologies of GDSC facilitate large-scale experiments on drug responses in cancer cell lines, ensuring a comprehensive understanding of genetic factors affecting drug metabolism and efficacy, thus enhancing reliability. In this paper, we only utilise the drug response (LN_IC50) data from GDSC.

2) Cancer Cell Line Encyclopedia (CCLE): The CCLE dataset, comprising gene expression data from 19221 genes across 1406 cancer cell lines representing 33 primary diseases, serves as a crucial resource obtained from the Broad Institute [37]. Widely recognised for its importance in anticancer drug response investigation, the CCLE dataset is instrumental in drug resistance model development and is often combined with GDSC drug response data [38, 39] due to the CCLE dataset's limited drug coverage, which may lead to poor generalisation of models despite providing a complete set of gene expression information. We leverage the CCLE gene expression data for model development, representing various types of cancer cell lines. This dataset is a vital resource for identifying cancer cells, as different cell types exhibit distinct gene expression levels.

3) PubChem: The PubChem database, comprising 118 million drug compounds and 318 million drug substances, is a vast resource for drug discovery, offering comprehensive chemical information freely accessible [40, 41]. It is widely utilised for bioactivity and toxicity prediction models, ligand discovery, and identifying new drug targets. We identified the Isosmiles feature in the PubChem dataset, which provides detailed molecular information crucial for understanding drug characteristics. However, Isosmiles data is in string format and unsuitable for direct model training. To convert Isosmiles data into a usable numerical format, we decided to integrate the Morgan Fingerprint technique [42]. It maps molecular structures into a binary representation by examining the radius of organic molecule bonds. This transformation allowed us to incorporate essential chemical compound information into our model, enhancing its ability to distinguish between drugs accurately.

B. Results

Throughout our model development process, we pursued two primary approaches: SVR and DNN approaches. Before initiating any hyperparameter tuning or feature selection, we established baseline models with standard parameters and assessed their predictive performance on the testing dataset. We opted to evaluate model performance using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearson Correlation (PC), and R-squared (R2) metrics. The evaluation results for the baseline DNN and SVR models are summarised in Table I. We also further refined and improved the DNN baseline model, as shown in Fig. 2, which exhibited superior performance compared to the initial models. We achieved an RMSE value of approximately 0.7, falling below 1, confirming the validity of our prediction results, meeting one of our model's requirements of having RMSE < 1.

To validate the performance and reliability of our model and to ensure that our results are comparable to other existing models, we conducted a benchmarking study against the results of the NeuPD [43]. Our benchmarking process utilised the same drug settings as detailed in Table II and Table III, although there were some deviations in our approach. Our model was trained specifically on breast cancer cell lines, whereas NeuPD did not focus on a specific cancer type. Furthermore, their study employed separate models for the GDSC and CCLE datasets, training each model on a distinct data source. In contrast, our approach integrated these data sources into a single cohesive dataset, combining the strengths of both for a more comprehensive training foundation. Another deviation was the dataset version; our model was trained using the updated GDSC2 dataset, while the NeuPD study used the GDSC1 dataset. Despite these differences, we maintained a similar dataset structure and settings to ensure comparability.

TABLE I. Developed Models Prediction Results using Testing Dataset (10% of the Drug-Breast-Chem Dataset, Random State = 42)

Model	Baseline DNN	Baseline SVR	Final DNN
RMSE	±1.330331	±2.494145	**±0.734142**
MAE	±0.996849	±1.931800	**±0.523844**
PC	±0.882805	±0.302953	**±0.956782**
R^2	±0.776131	±0.020114	**±0.915103**

TABLE II. Benchmarking Setting 1 (GDSC Case) - Drugs Used

Drug Name	PubChem CID
Afatinib	10184653
Alectinib	49806720
Bleomycin (50 uM)	5460769
Dabrafenib	44462760
GNF-2	5311510
Linifanib	11485656
Methotrexate	126941
Quizartinib	24889392
SN-38	104842
Trametinib	11707110

TABLE III. Benchmarking Setting 2 (CCLE Case) - Drugs Used

Drug Name	PubChem CID	Drug Name	PubChem CID
AEW541	11476171	Irinotecan	60838
Nilotinib	644241	Topotecan	60700
17-AAG	6505803	LBW242	11503417
PHA-665752	10461815	PD-0325901	9826528
Lapatinib	208908	PD-0332991	5330286
Nutlin-3	216345	Paclitaxel	36314
AZD0530	10302451	AZD6244	10127622
PF2341066	11626560	PLX4720	24180719
L-685458	5479543	RAF265	11656518
ZD-6474	3081361	TAE684	16038120
Panobinostat	6918837	TKI258	135611162
Sorafenib	216239	Erlotinib	176870

For the benchmarking, we preprocessed the dataset to create two datasets that included only the drugs specified in the settings outlined in Tables IV and Table V. This preprocessing ensured that our benchmarking settings were as aligned as possible with those used in the NeuPD study, even though slight deviations occur due to different versions of GDSC dataset and different focus on cell lines. By utilising these settings, we conducted two sets of tests using both the baseline SVR model and our final DNN model. The benchmarking results, presented in Table IV, provide a comprehensive comparison of our model's performance against the models reported in the NeuPD study. The results show that our final DNN model achieved superior benchmarking performance compared to the baseline SVR model and demonstrated comparable or better performance relative to the models of conventional machine learning approaches reported in the NeuPD study. The DNN model's lower RMSE values indicate higher accuracy and robustness, validating our approach and the improvements made over conventional machine learning methods. However, upon comparing our final DNN model with the NeuPD model, which also utilised deep learning approaches, we found that the NeuPD model exhibited overall better performance. This discrepancy may be attributed to slight deviations in the benchmarking dataset, including information loss due to GDSC data versioning and different data combination approaches. The NeuPD model demonstrated strong predictive capabilities, particularly with the GDSC dataset. This finding supports our decision to use a combined GDSC-CCLE dataset due to our model's better performance in the CCLE settings. Despite these differences, our model still delivered valid predictions with satisfactory performance, proving to be competitive with the NeuPD model.

We also examined two other models: DeepDSC [35] and DeepCDR [44]. NeuPD [43] has provided a state-of-the-art comparison of these models' performances against other existing models. Table V below shows the overall comparisons, further confirming that our model produced valid predictions, indicating great achievement. Upon further comparison with the model utilising the GDSC-CCLE combined dataset approach, namely the DeepCDR model, our proposed DNN model demonstrates notable improvements. While the DeepCDR model achieves an RMSE of approximately 1.058, our model achieves a lower RMSE of around 0.75. It is important to note that the DeepCDR model conducts predictions across multiple TCGA cancer types, whereas our focus is solely on breast cancer. Despite this distinction, our model's RMSE reflects a significant improvement compared to DeepCDR, showcasing enhanced predictive accuracy within the context of breast cancer drug resistance prediction.

TABLE IV. Benchmarking Results Achieved

Setting 1 (GDSC case for NeuPD)				
Method	**RMSE**	**MSE**	**MAE**	**R^2**
ElasticNet	±2.419	±5.851	±2.020	0.532
SVR	**±2.104**	**-**	**±1.882**	**±0.093**
XGBoost	±1.337	± 1.794	±0.953	0.609
DNN	**±0.713**	**-**	**±0.542**	**±0.893**
NeuPD	±0.490	±0.246	±0.392	0.929
Setting 2 (CCLE case for NeuPD)				
Method	**RMSE**	**MSE**	**MAE**	**R^2**
ElasticNet	±3.202	±10.25	±2.972	0.281
SVR	**±2.514**	**-**	**±1.579**	**±0.132**
XGBoost	±2.074	±4.308	±1.491	0.317
NeuPD	±1.784	±3.192	±1.568	0.543

DNN	**±1.108**	**-**	**±0.998**	**±0.841**

TABLE V. ACHIEVED RMSE UPON COMPARISONS WITH OTHER EXISTING MODELS

Setting 2 (CCLE case)	Ridge Regression [45]	±6.576
	Random Forest [45]	±5.738
	Elastic Net [45]	±5.378
	Lasso [45]	±5.333
	ElasticNet [43]	±3.202
	SVR	±2.514
	XGBoost [43]	±2.074
	NeuPD [43]	±1.784
	DNN	**±1.108**

C. Discussions

Overall, the predictive performance of the DNN model can be considered highly effective and valid based on the obtained evaluation metrics. The RMSE of 0.7341, representing the average deviation of predicted LN_IC50 values from the actual values, showcases the model's accuracy. With this RMSE value, the model's predictions typically deviate by approximately 0.7341 units from the true LN_IC50 values. This level of accuracy (loss) is commendable, especially given the complexity of drug response prediction in breast cancer cell lines. The relatively low RMSE suggests that the model has successfully captured the underlying patterns in the data, making it a reliable tool for predicting drug sensitivity.

The MAE of 0.5238 further supports the model's robustness. MAE, being a straightforward measure of prediction error, indicates that the model's predictions deviate by about 0.5238 units from the actual values on average. This low MAE demonstrates the model's ability to consistently make accurate predictions across different data points. The smaller the MAE, the better the model is at making precise predictions, and in this case, the low MAE reinforces the model's effectiveness in predicting LN_IC50 values accurately.

The Pearson Correlation coefficient of 0.9568 is particularly noteworthy. This high value, close to 1, indicates a very strong positive linear relationship between the predicted and actual LN_IC50 values. Such a strong correlation suggests that the model's predictions are not only accurate but also consistently aligned with the actual data trends. This level of correlation is significant because it implies that the model has captured the inherent relationships within the dataset, which is critical for making reliable predictions in real-world applications.

The R2 value of 0.9151 further underscores the model's high performance. An R2 value above 0.9 indicates that the model explains over 91% of the variance in the actual LN_IC50 values. This high explanatory power is crucial for confidence in the model's predictions and suggests that the model has effectively learned from the data. The high R2 value, combined with the other metrics, reflects the model's capability to generalise well to new, unseen data.

To further validate our model's predictive capability, we generated a random testing dataset of 288 rows and utilised our model to predict LN_IC50 values. Subsequently, we applied a resistance cut-off to classify drugs as sensitive or resistant. Our aim was to assess the model's ability to correctly identify drug sensitivity, which is crucial for effective cancer treatment. Through classification testing, we evaluated the model's performance in distinguishing between sensitive and resistant drugs. This classification is crucial for determining suitable cancer treatments avoiding ineffective drugs with high resistance levels. The confusion matrix in Fig. 4 illustrates the results of this evaluation, with True Positive (TP) representing correctly identified low-resistance drugs, True Negative (TN) indicating accurately identified high-resistance drugs, False Positive (FP) denoting falsely identified low-resistance drugs, and False Negative (FN) representing falsely identified high-resistance drugs.

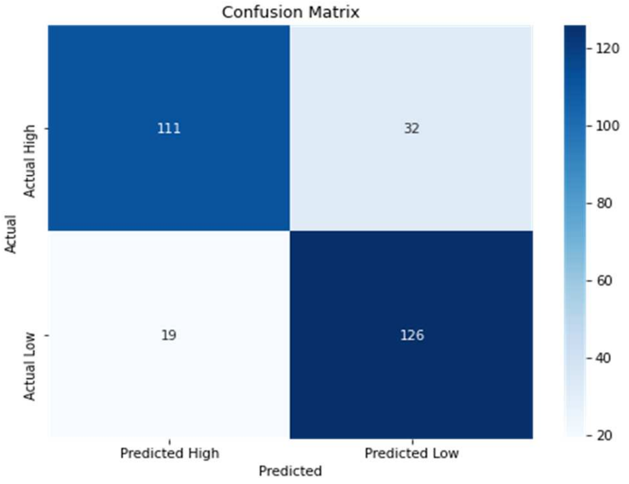


Fig. 4. Confusion Matrix of Prediction Result on Random Testing Dataset

The confusion matrix yielded 126 true positives, indicating correct identification of low-resistance drug-cell pairs, 111 true negatives, indicating accurate identification of high-resistance pairs, and 32 false positives and 19 false negatives, representing misclassifications of resistance levels. Despite some misclassifications, the overall accuracy, recorded at 0.82 and precision at [0.80,0.85], suggests that our model performs reasonably well in identifying drug resistance levels, as the majority of the predicted LN_IC50 values closely match their actual resistance levels.

The precision values further underscore the model's effectiveness. Precision for low-resistance predictions is 0.80, and for high-resistance predictions, it is 0.85. Precision measures the proportion of true positive predictions among all positive predictions, indicating how often the model's positive predictions are correct. In this context, a precision of 0.80 means that 80% of the drugs predicted to have low resistance indeed have low resistance, while a precision of 0.85 means that 85% of the drugs predicted to have high resistance truly exhibit high resistance. These values are significant as they reflect the model's accuracy in identifying effective (low-resistance) and ineffective (high-resistance) drugs, thereby reducing the risk of false positives and ensuring more reliable treatment plans.

However, further refinement may be necessary to improve the accuracy and precision of sensitivity classification. Overall, these findings provide valuable insights into the effectiveness of our predictive model and its potential application in guiding cancer treatment decisions, and it is proven that our model has successfully achieved a valid prediction outcome.

V. Conclusion

In summary, this paper has successfully developed a comprehensive and valid integrated dataset for the data preprocessing phase, ensuring it contained sufficient cancer cell and drug information, including gene expression data from CCLE, drug response data from GDSC, and drug chemical information from PubCHEM. The successful application of the Morgan Fingerprint technique to the drug's chemical compound features ensured effective integration into the model for training. For the model development phase, we developed two models using DNN and SVR approaches. Ultimately, we focused on the DNN approach due to its superior prediction capability and made significant improvements. Our model's performance surpassed some existing works and conventional machine learning approaches reviewed, such as NeuPD, DeepDSC, and DeepCDR. We achieved a prediction RMSE below 1, meeting our expectations for valid predictions.

Currently, our predictive model is tailored specifically for breast cancer, with the training data exclusively comprising breast cancer cell lines. To broaden the applicability of our model for various cancer treatments, it is essential to extend our training datasets to include multiple cancer types in future works. This can be achieved by developing separate models for different cancer types, each trained on Gene-Drug-Chem datasets specific to those types. An ensemble approach could be particularly promising, where individual models for different cancer types are combined.

Beyond validating predictive performance, our model's results suggest significant potential for personalised cancer treatment, allowing clinicians to tailor therapies based on individual patient characteristics and enhance treatment efficacy. Ongoing refinement and optimisation are crucial for boosting predictive accuracy and clinical utility. Future research could integrate additional features or data sources to address misclassifications in sensitivity classification. Validation studies using independent datasets and real-world clinical outcomes will provide valuable insights into the model's generalisability and applicability, laying a solid foundation for future advancements in precision oncology and personalised medicine.

Acknowledgements

This work is supported by the Ministry of Higher Education (MOHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS) (Ref: FRGS/1/2022/ICT06/MUSM/03/1).

References

[1] "Cancer." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer (accessed Sep. 23, 2023).

[2] F. S. Collins, E. D. Green, A. E. Guttmacher, M. S. Guyer, and U. S. N. H. G. R. I. on behalf of the, "A vision for the future of genomics research," Nature, vol. 422, no. 6934, pp. 835-847, 2003/04/01 2003, doi: 10.1038/nature01626.

[3] K. Zaitsev, M. Bambouskova, A. Swain, and M. N. Artyomov, "Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures," Nature Communications, vol. 10, no. 1, p. 2209, 2019/05/17 2019, doi: 10.1038/s41467-019-09990-5.

[4] H. Yuan, Q. Ma, L. Ye, and G. Piao, "The Traditional Medicine and Modern Medicine from Natural Products," Molecules, vol. 21, no. 5, p. 559, 2016. [Online]. Available: https://www.mdpi.com/1420-3049/21/5/559.

[5] K. Do et al., "Phase I Study of Single-Agent AZD1775 (MK-1775), a Wee1 Kinase Inhibitor, in Patients With Refractory Solid Tumors," Journal of Clinical Oncology, vol. 33, no. 30, pp. 3409-3415, 2015, doi: 10.1200/jco.2014.60.4009.

[6] A. Kalamara, L. Tobalina, and J. Saez-Rodriguez, "How to find the right drug for each patient? Advances and challenges in pharmacogenomics," Current Opinion in Systems Biology, vol. 10, pp. 53-62, 2018/08/01/ 2018, doi: https://doi.org/10.1016/j.coisb.2018.07.001.

[7] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," Briefings in Bioinformatics, vol. 22, no. 1, pp. 360-379, 2020, doi: 10.1093/bib/bbz171.

[8] J. Lv, G. Liu, Y. Ju, Y. Sun, and W. Guo, "Prediction of Synergistic Antibiotic Combinations by Graph Learning," Frontiers in Pharmacology, vol. 13, p. 849006, 03/01 2022, doi: 10.3389/fphar.2022.849006.

[9] S. Liang and H. Yu, "Revealing new therapeutic opportunities through drug target prediction: a class imbalance-tolerant machine learning approach," Bioinformatics, vol. 36, no. 16, pp. 4490-4497, 2020, doi: 10.1093/bioinformatics/btaa495.

[10] A. Zhang, K. Miao, H. Sun, and C. X. Deng, "Tumor heterogeneity reshapes the tumor microenvironment to influence drug resistance," (in eng), Int J Biol Sci, vol. 18, no. 7, pp. 3019-3033, 2022, doi: 10.7150/ijbs.72534.

[11] T. D. Laajala, T. Gerke, S. Tyekucheva, and J. C. Costello, "Modeling genetic heterogeneity of drug response and resistance in cancer," (in eng), Curr Opin Syst Biol, vol. 17, pp. 8-14, Oct 2019, doi: 10.1016/j.coisb.2019.09.003.

[12] N. Billows, J. E. Phelan, D. Xia, Y. Peng, T. G. Clark, and Y. M. Chang, "Feature weighted models to address lineage dependency in drug-resistance prediction from Mycobacterium tuberculosis genome sequences," (in eng), Bioinformatics, vol. 39, no. 7, Jul 1 2023, doi: 10.1093/bioinformatics/btad428.

[13] G. Housman et al., "Drug Resistance in Cancer: An Overview," Cancers, vol. 6, no. 3, pp. 1769-1792, 2014. [Online]. Available: https://www.mdpi.com/2072-6694/6/3/1769.

[14] T. Haider, V. Pandey, N. Banjare, P. N. Gupta, and V. Soni, "Drug resistance in cancer: mechanisms and tackling strategies," Pharmacological Reports, vol. 72, no. 5, pp. 1125-1151, 2020/10/01 2020, doi: 10.1007/s43440-020-00138-7.

[15] F. S. Collins, "Medical and Societal Consequences of the Human Genome Project," New England Journal of Medicine, vol. 341, no. 1, pp. 28-37, 1999, doi: doi:10.1056/NEJM199907013410106.

[16] S. Chawla et al., "Gene expression based inference of cancer drug sensitivity," Nature Communications, vol. 13, no. 1, p. 5680, 2022/09/27 2022, doi: 10.1038/s41467-022-33291-z.

[17] W. Yang et al., "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," Nucleic Acids Research, vol. 41, no. D1, pp. D955-D961, 2012, doi: 10.1093/nar/gks1111.

[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, 2012/03/01 2012, doi: 10.1038/nature11005.

[19] C. Huang, R. Mezencev, J. F. McDonald, and F. Vannberg, "Open source machine-learning algorithms for the prediction of optimal cancer drug therapies," PLOS ONE, vol. 12, no. 10, p. e0186906, 2017, doi: 10.1371/journal.pone.0186906.

[20] J. C. Costello et al., "A community effort to assess and improve drug sensitivity prediction algorithms," Nature Biotechnology, vol. 32, no. 12, pp. 1202-1212, 2014/12/01 2014, doi: 10.1038/nbt.2877.

[21] M. Gönen and A. A. Margolin, "Drug susceptibility prediction against a panel of drugs using kernelised Bayesian multitask learning," Bioinformatics, vol. 30, no. 17, pp. i556-i563, 2014, doi: 10.1093/bioinformatics/btu464.

[22] I. Cortés-Ciriano et al., "Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel," Bioinformatics, vol. 32, no. 1, pp. 85-95, 2015, doi: 10.1093/bioinformatics/btv529.

[23] S. Naulaerts, C. C. Dang, and P. J. Ballester, "Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours," Oncotarget, vol. 8, no. 57, 2017. [Online]. Available: https://www.oncotarget.com/article/20923/text/.

[24] K. M. Gayvert, O. Aly, J. Platt, M. W. Bosenberg, D. F. Stern, and O. Elemento, "A Computational Approach for Identifying Synergistic Drug Combinations," PLOS Computational Biology, vol. 13, no. 1, p. e1005308, 2017, doi: 10.1371/journal.pcbi.1005308.

[25] M. P. Menden et al., "Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties," PLOS ONE, vol. 8, no. 4, p. e61318, 2013, doi: 10.1371/journal.pone.0061318.

[26] Z. Dong et al., “Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection,” BMC Cancer, vol. 15, no. 1, p. 489, 2015/06/30 2015, doi: 10.1186/s12885-015-1492-6.

[27] M. Ammad-ud-din, S. A. Khan, K. Wennerberg, and T. Aittokallio, “Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression,” Bioinformatics, vol. 33, no. 14, pp. i359-i368, 2017, doi: 10.1093/bioinformatics/btx266.

[28] X. Xu, H. Gu, Y. Wang, J. Wang, and P. Qin, "Autoencoder Based Feature Selection Method for Classification of Anticancer Drug Response," (in eng), Front Genet, vol. 10, p. 233, 2019, doi: 10.3389/fgene.2019.00233.

[29] K. Koras, D. Juraeva, J. Kreis, J. Mazur, E. Staub, and E. Szczurek, “Feature selection strategies for drug sensitivity prediction,” Scientific Reports, vol. 10, no. 1, p. 9377, 2020/06/10 2020, doi: 10.1038/s41598-020-65927-9.

[30] E. Oren et al., “Pan-genome and multi-parental framework for high-resolution trait dissection in melon (Cucumis melo),” The Plant Journal, vol. 112, no. 6, pp. 1525-1542, 2022, doi: https://doi.org/10.1111/tpj.16021.

[31] H.-L. Her and Y.-W. Wu, “A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains,” Bioinformatics, vol. 34, no. 13, pp. i89-i95, 2018, doi: 10.1093/bioinformatics/bty276.

[32] N. Batis, J. M. Brooks, K. Payne, N. Sharma, P. Nankivell, and H. Mehanna, “Lack of predictive tools for conventional and targeted cancer therapy: Barriers to biomarker development and clinical translation,” Advanced Drug Delivery Reviews, vol. 176, p. 113854, 2021/09/01/ 2021, doi: https://doi.org/10.1016/j.addr.2021.113854.

[33] M. P. Menden et al., “Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen,” Nature Communications, vol. 10, no. 1, p. 2674, 2019/06/17 2019, doi: 10.1038/s41467-019-09799-2.

[34] B. M. Kuenzi et al., “Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells,” Cancer Cell, vol. 38, no. 5, pp. 672-684.e6, 2020, doi: 10.1016/j.ccell.2020.09.014.

[35] M. Li et al., “DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 2, pp. 575-582, 2021, doi: 10.1109/TCBB.2019.2919581.

[36] Y. Shen and Z. Yan, “Systematic prediction of drug resistance caused by transporter genes in cancer cells,” Scientific Reports, vol. 11, no. 1, p. 7400, 2021/04/01 2021, doi: 10.1038/s41598-021-86921-9.

[37] J. Barretina et al., “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity,” Nature, vol. 483, no. 7391, pp. 603-607, 2012/03/01 2012, doi: 10.1038/nature11003.

[38] D. Wei, C. Liu, X. Zheng, and Y. Li, “Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model,” BMC Bioinformatics, vol. 20, no. 1, p. 44, 2019/01/22 2019, doi: 10.1186/s12859-019-2608-9.

[39] F. Xia et al., “A cross-study analysis of drug response prediction in cancer cell lines,” Briefings in Bioinformatics, vol. 23, no. 1, 2021, doi: 10.1093/bib/bbab356.

[40] S. Kim, “Getting the most out of PubChem for virtual screening,” Expert Opinion on Drug Discovery, vol. 11, no. 9, pp. 843-855, 2016/09/01 2016, doi: 10.1080/17460441.2016.1216967.

[41] S. Kim et al., “PubChem Substance and Compound databases,” Nucleic Acids Research, vol. 44, no. D1, pp. D1202-D1213, 2015, doi: 10.1093/nar/gkv951.

[42] D. Medin. “Data Science for drug discovery research -Morgan fingerprints in Python.” Medium. (accessed Jan. 2024).

[43] M. Shahzad, M. A. Tahir, M. Alhussein, A. Mobin, R. A. Shams Malick, and M. S. Anwar, “NeuPD—A Neural Network-Based Approach to Predict Antineoplastic Drug Response,” Diagnostics, vol. 13, no. 12, p. 2043, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/12/2043.

[44] Q. Liu, Z. Hu, R. Jiang, and M. Zhou, “DeepCDR: a hybrid graph convolutional network for predicting cancer drug response,” Bioinformatics, vol. 36, no. Supplement_2, pp. i911-i918, 2020, doi: 10.1093/bioinformatics/btaa822.

[45] Q. Li, R. Shi, and F. Liang, “Drug sensitivity prediction with high-dimensional mixture regression,” PLOS ONE, vol. 14, no. 2, p. e0212108, 2019, doi: 10.1371/journal.pone.0212108