Explainable AI in Drug Sensitivity Prediction on Cancer Cell Lines

Ismat Saira Gillani ,[1] **Dr. Muhammad Shahzad**[2]

Ansharah Mobin[3] **Muhammad Rizwan Munawar**[4]

Muhammad Usman Awan[5] **Muhammad Asif**[6]

1 ABSTRACT

Explainable Artificial Intelligence (XAI) is a field that develops ways to explain predictions made by AI models. In this paper XAI which is a multifaceted approach is discussed which is capable of defining the value of features while producing predictions. Precision medicine and the forecast of cancer's reaction to a specific treatment or drug efficiency is an area of active research. Drug sensitivity forecasting on massive genomics data is a strenuous process in drug discovery. However, drug personalization on the other hand is a tedious and arduous matter. Explainable AI is one of the many properties that instills confidence and dependency in AI systems which is why more attention needs to be paid to XAI. This research is a step toward a more profound understanding of deep learning techniques [1] on gene expressions and drug chemical structures.

Keywords: Drug sensitivity, Drug similarity, Cell lines, Explainable AI, Personalized drugs.

2 INTRODUCTION

Since cancer is a common human genetic disease caused due to irregular growth of human cells. The human genetic system is so complex that makes it really challenging to treat cancer. There is no universal medicine that works for every patient, each patient responds to the drug in a different way. Personalized medicine uses human genomic profiles or proteins to thwart and diagnose disease. Genomic information is used to observe an individual's responses to drugs. If a gene variant is associated with specific drug response in a patient, doctors then make a decision based on genetics by adjusting the prescribed amount of drug or picking a different drug. The complex mechanism of drug action and the high heterogeneity of cancer lessens the response rate of most anti-cancer drugs. Drug sensitivity states the concept that the bacteria cannot be produced if the drug is present and demonstrates that the drug or antibiotic is effective for those bacteria or cell lines[1]. In this research, XAI is discussed as a method that can be used in the diagnosis and analysis of drugs. The proposed approach is presented with the intention of attaining transparency, accountability, and model improvement in drug sensitivity prediction on cancer cell lines. In particular, Explainable Artificial Intelligence (XAI) has been recognized as a practical working method for concluding the relevance of important features when making predictions using Machine Learning (ML) models having high local fidelity. Thus XAI findings could lead to accurate clinical predictions. The response of the drugs being different is factorized into signaling pathway drug target feature [2]. It is an important measure for determining the drug as being sensitive or resistant. The association between drugs and human ge-

nomics profile is unveiled by executing HTS-High throughput screening and is open-sourced, and accessible in the form of pharmacogenomics datasets like Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE). In a dataset of Cancer Cell Line Encyclopedia (CCLE), the compilation of chromosomal profiles of 947 human cancer cell lines and profile of 24 anticancer drugs around 500 cancel lines to support personalized medicine for cancer patients. What makes it extra helpful in drug response prediction is that the cell line in CCLE and Genomics of Drug Sensitivity in Cancer (GDSC) datasets have acquired the data from various types of cancer tissues. Some of which are lungs, brain, kidney, breast. The main idea is to predict the drug responses of cancer cell lines using GDSC and CCLE datasets and assess the drug as being sensitive or unaffected in terms of IC50 value. In our research, we modeled an architecture inspired by deep learning for drug response motivated by the expeditious development of deep learning technology. The prediction of cancer cell lines is achieved by incorporating chemical profiles of compounds and genomic profiles of cell lines. The approach involved the prediction of the response values IC50, the model comprises a neural network, which worked in an unsupervised way to extract cell lines features from gene expression data. The dimension of gene expression data is extremely massive, hence the Pearson correlation was used to reduce dimensions. Afterward, to make drug sensitivity data of the particular cell line-compound pairs, the chemical features of compounds were incorporated into the model. For evaluation 10-fold cross-validation technique was used. Coefficient of determination and Root Mean Square Error (RMSE) were the loss functions. The Shapley values were calculated which can evaluate feature importance for a specific prediction for any model. However, XAI is a niche focused on enhancing the explainability and clarity of AI algorithms. The popular XAI methods which include 'Local Interpretable Model-Agnostic Explanations' (LIME) and 'SHAPely Additive Explanations' (SHAP) have proven fruitful in the interpretation of black box models. A lot of methods have been proposed in recent studies, such as modular relations among genomics features and models built on deep neural networks [3]. In this research, first the gaps were investigated. We then came up with a more improved algorithm that can efficiently address the gaps of earlier research. The main focus was to translate critical information which was produced during prediction so that personalized medicine can be further improved and medical practitioners can use that translatable information [4] more effectively. Moreover, the model reveals that deep learning [5] can significantly promote the study of drug sensitivity prediction. To learn the molecular basis of drug sensitivity [6], gene expressions and cancer cell lines with diverse genomics background is studied. Multiple concepts of AI have also been effectively implemented for personalized drug discoveries in earlier years. Quite extensive research has been done on these data sets, for instance, anti-cancer drug repositioning [7] Cancer pathogenesis analysis [8] etc. Though, extensive experiments have proved that modeling cell line features alongside chemical features may bring great results for drug sensitivity predictions. In this literature review, the majority of the discussed approaches have predicted drug response by using machine learning algorithms. In [9] Aman proposed a method for mapping non-linear interconnection among the response of drugs and gene expressions of cell lines by utilizing multi-task and ensemble learning. In [10] ensemble technique was suggested. Which was built on rotation forest, which is a technique based on feature extraction for producing classifier ensembles.[11] Liu's combination model accomplished higher prediction accuracy as well as great interpretation ability. The ensemble learning method by Mehmat [12] gave promising results. He also proposed drug activity signatures and cell line sensitivity by mixing multiple databases and observed its effect on the performance of the response variable. As compared to traditional machine learning algorithms of SVM and Random Forest [13] their model accomplished significantly improved performance.

Even though the DL model is effectively used in numerous applications, the use of large and complex datasets restricts its use in integrating genomic feature sets to predict the drug response. To handle this challenge [14] suggested a deep neural network model built on mutation information and the expression profile of the cancer cells. The model included three networks (mutation encoder, expression, and feed-forward network), where mutation encoders and the expressions were mutually responsible for dimensionality reduction. The reduced dataset was then put into the FFNN for predicting drug response based on the IC50 values. Towards Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-based Convolutional Encoders [15] They proposed the architecture of anticancer compound sensitivity for explainable prediction by utilizing a multimodal attention-based convolutional encoder. [15] In this study, the authors have worked on presenting a novel framework for patient classification and drug reprofiling by using exploratory data mining and network analysis. This is a step toward building advanced Explainable AI systems. Manica, Ali [16] AI Enables Explainable Drug Sensitivity Screenings. The IBM Researchers established PacMan an insilico stage for screening compounds which is based on the latest developments in AI for computational biochemistry. In their findings, the genes with the highest attention weights were the main participants in the development and treatment of the disease.

3 METHODS AND TECHNIQUES

In this research, a deep learning model is used for predicting drug sensitivity on cancer cell lines. The use of compound fingerprints and cell line expression data as an input for the model. All procedures are implemented using Python and its libraries scikit learn, tensor flow and keras. RDKit functionality from python was used to find the hash values of Morgan fingerprints on the drugs. Pearson correlation technique was used to reduce the dimensions of data to the fortieth of the original data i.e. to 500 features. Morgan fingerprints Fig 3, of the 1D and 2D compound structure and drug chemical features were put into the model. To deal with the problem of gradient explosion, the gradient absolute value is clipped to keep its value under five. Four layered feed-forward neural network with 756 neural units in their input layer was used. 756 neurons were the result count of 500 extracted relevant features of cell lines and the length of the Morgan fingerprints i.e., 256. The first layer of the neural network has 1000 neurons, the second layer has 800, the third layer has 500 and the fourth layer has 100 neurons. ELU is the activation function in four hidden layers. RMSE and MAE are the loss function. Early stopping was at 30. The dropout rate of 0.1 is used after three layers. The learning rate was 0.0004, gradient explosion problem was handled by clipping the gradient absolute value in the range of 5.

3.1 Data Gathering

For training of the model, GDSC and CCLE datasets were used which are widely accessible on GDSC and CCLE information banks. Drug sensitivity data of 24 drugs and 491 cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) were downloaded via the DepMap portal. Additionally, the SDF values for each drug are taken out from Pub- Chem. The strings of Canonical SMILE were gathered by finding the Pub Chem database with an open-source Python API, PubChemPy. It transformed SMILE strings into Morgan fingerprints by using RDKit, which is a cheminformatics toolkit, and generated the matrix of 24 drugs by 256 molecular bits for the CHEM (chemical structure feature). In the Gene expression dataset, rows represent genes and columns represent cell lines, values of genes in cell lines represent the activation time when the gene became effective. The IC50 values were transformed into -log (IC50). higher IC50 means lower values in -log (IC50).

Table 1. Drug sensitivity analysis using GDSC and CCLE datasets

Datasets	Cell lines	Drugs
GDSC	655	139
CCLE	491	24

3.2 Data Preprocessing

The suggested method includes the following steps. GDSC and CCLE data sets were taken, missing values were imputed with the mean values. Data was normalized and brought in the range of 0 to 1. The dataset being highly dimensional, numerous dimensionality reduction techniques were used but Pearson accomplished promising results. 500 features were extracted. These 500 features were then concatenated with 256 Morgan fingerprints data. That final set of data with 756 features was fed into Neural Network.

3.3 Illustration of the method

The flow chart of the model shows all the steps taken during training. X denotes the input and "h" denotes the 500 relevant features, which were then combined with Morgan fingerprints and produced the matrix with the dimension of 756. Data was then fed into Feedforward neural network. Shapely values were calculated of GDSC and CCLE, top 20 features from both data sets were then analyzed to see if there are any common features picked out but SHAPley algorithm.

3.4 External Validation

GDSC dataset was used for training and CCLE data set was used for testing.

4 EVALUATION MEASURES

In the evaluation phase, a 10-fold cross-validation technique was conducted to validate the performance of our model in terms of root-mean-square error (RMSE) which is the most useful technique when big errors are particularly undesirable. It is observed that our model outperformed the previous approaches with RMSE of 0.52 on GDSC dataset, and

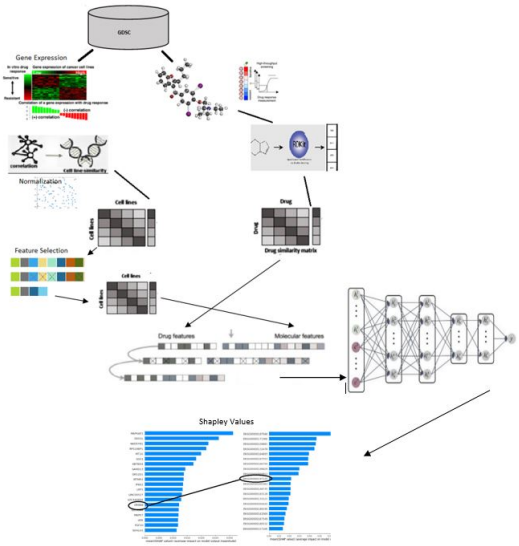

Fig. 1. Flowchart

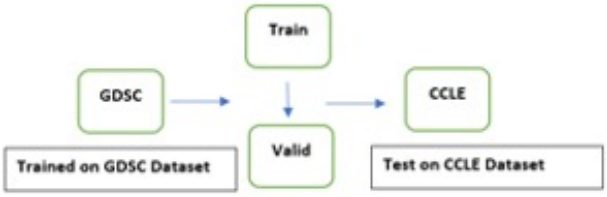

Fig. 2. External Validation using GDSC and CCLE

RMSE of 0.23 on CCLE dataset. The R2 and RMSE of tenfold cross-validation were obtained as shown in the table below.

Table 2.

Neural Network	RMSE	R^2
GDSC	0.52	0.78
CCLE	0.23	0.78

4.1 Comparison with Other Methodologies

The proposed approach showed amazing results when compared with other methodologies, whereas PathDSP performed well. State-of-the-art results are expected in the future when different combinations of genomic profiles including copy number variation, pathways, mutations, and drug-target interactions will be part of training data.

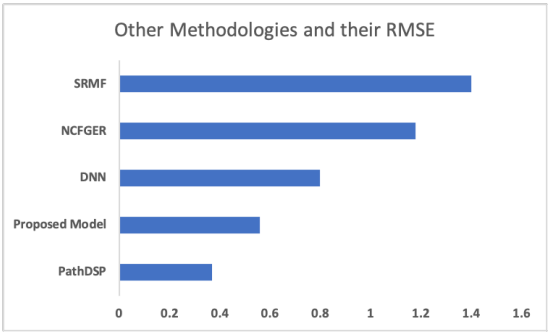

Fig. 3. Results obtained.

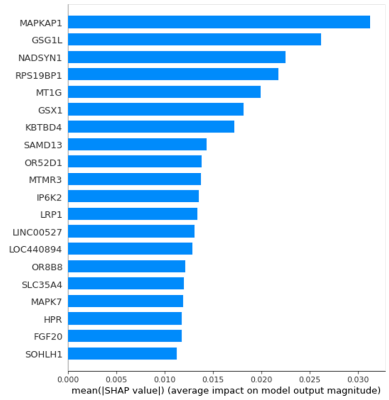

Fig. 4. CCLE Dataset- Mean Shap values on the x-axis and cell lines on the y-axis.

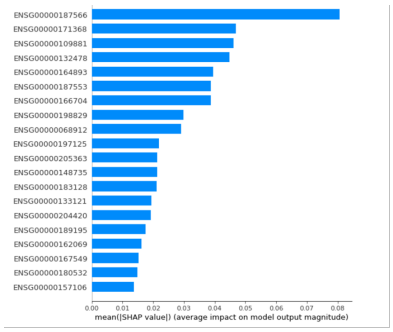

Fig. 5. GDSC Dataset- Mean Shap values on the x-axis and cell lines on the y-axis.

5 EXPLAINABLE AI - XAI

The aim was to address the issue of the unexplained ability of features of any deep learning/machine learning model. The objective was therefore achieved by using Explainable AI, XAI is a framework and method which can interpret and recognize predictions made by the model. There are several ways in which predicted values of the model can be translated. Here the SHAP values were used.

5.1 Computed Shapley values to identify important features

It is a technique of describing the significance of predictive features by assigning them scores. These scores indicate the relative importance of each feature when making a prediction, in other words, the amount of contribution made by an individual feature makes to the prediction value. In this research Shapely Values were used, these values were calculated by python library SHAP, Shapley values recognize significant features under-lying predictions. Based on shapely values further analysis will be done in the future as well.

5.2 Prediction Explanation

Quantification of feature importance by using Shapley values. Below are the shapely values calculated on GDSC and CCLE data sets.

Shap value represents the contribution of that particular cell line in predicting drug sensitivity. The below figure shows that the drug target OR8B8/ENSG00000197125 in GDSC as well as in the CCLE data set has a positive contribution to drug sensitivity prediction. Based on this drug target and other drug targets with positive contributions to drug sensitivity prediction, new drugs can be designed or it can also be said that while designing new anti-cancer drugs these critical drug targets should be given utmost importance as these have high contributions in causing cancer.

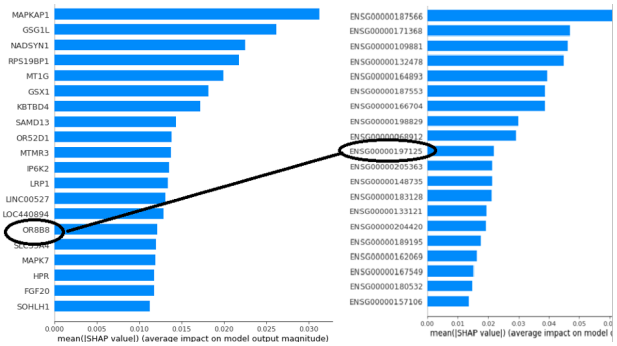

Fig. 6. Drug Target

6 CONCLUSION

The proposed techniques based on deep neural networks for drug sensitivity prediction have shown improved performance. In addition to this, it has made the predictions explainable. In the drug discovery scenario, high accuracy of DL models is not achievable, but the contributed predictions can be of great value to doctors. XAI [16] is expected to provide vital support in the analysis and interpretation of complex data. In the Future, it can be used on a different combination of genomic profiles which includes copy number variation, pathways, mutations, and drug-target interactions. The combination of genomic profiles can help in making a cell line similarity matrix with a high correlation which can help to better prediction. Currently, a single deep learning algorithm is used for prediction. Combining several algorithms can lead to precise results and the addition of an explainable AI can play a vital role in analysis and its interpretation.

References

[1] DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. Min Li, Yake Wang, Ruiqing Zheng, Xinghua Shi.

[2] Heming Zhang, Yixin Chen and Fuhai Li, "Predicting Anticancer Drug Response With Deep Learning Con- strained by Signaling Pathways" in Bioinform., 2021.

[3] Yu-Chiao Chiu, Hung-I Harry Chen, Tinghe Zhang, Songyao Zhang, Aparna Gorthi, Yufei Huang, Yidong Chen, "Predicting drug response of tumors from integrated genomic profiles by deep neural networks," BMC Medical Genomics, vol. 12, 2019.

[4] Jose´ Jime´nez-Luna, Drug discovery with explainable artificial intelligence.

[5] Ali Oskooei, Matteo Manica, Roland Mathis, Maria Rodriguez Martinez, Network-based Biased Tree En- sembles (NetBiTE) for Drug Sensitivity Prediction and Drug Sensitivity Biomarker Identification in Cancer. Sci Rep 9, 15918 (2019)

[6] Hui Liu, Yan Zhao, Lin Zhang, XingChen, "Anti- cancer Drug Response Prediction Using Neighbor- Based Collaborative Filtering with Global Effect Re- moval," Molecular Therapy - Nucleic Acids, vol. 13, pp. 303-311, 2018.

[7] ZainabAl-Taie, Explainable artificial intelligence in high-throughput drug repositioning for subgroup strat- ifications with intervention able potential.

[8] M. J. Garnett, E. J. Edelman, S. J. Heidorn et al., "Sys- tematic identification of genomic markers of drug sen- sitivity in cancer cells," Nature, vol. 483, no. 7391, pp. 570-5, Mar 28, 2012.

[9] Sharma Aman, Rani Rinkle, "KSRMF: Kernelized sim- ilarity based regularized matrix factorization frame- work for predicting anti-cancer drug responses," J. Intell. Fuzzy Syst., vol. 35, pp. 1779- 1790, 2018.

[10] Sharma, A., Rani, R. Drug sensitivity prediction frame- work using ensemble and multi-task learning. Int. J.Mach. Learn. Cyber. 11, 1231–1240 (2020).

[11] Liu, Chuanying Wei, Dong Xiang, Ju Ren, Fuquan Huang, Li Lang, Jidong Tian, Geng Li, Yushuang Yang, Jialiang. (2020). Improved Anti- cancer Drug Response Prediction Based on An En- semble, Method Integrating Matrix Completion and Ridge Regression. Molecular Therapy - Nucleic Acids. 21.10.1016/j.omtn.2020.07.003.

[12] Mehmet Tan,"Drug response prediction by ensemble learning and drug-induced gene expression signatures", Genomics, Volume 111, Issue 5, September 2019, Pages 1078-1088.

[13] Raziur Rahman "Heterogeneity Aware Random Forest for Drug Sensitivity Prediction", Scientific Reports volume 7, Article number: 11347 (2017).