Evaluating the Effectiveness of LightGBM for Drug Sensitivity to Cancer Line Prediction

Claudio Januar Aristan[a], Aloysius Kang[a], Vincent Dava Sutomo[a], Neo Cenon[a], Nesti F. Sianipar[b]

[a] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[b] *Lecturer, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Cancer comes in many types and can be located in different parts of the body. To treat each type of cancer, a long process of drug analysis is needed. For this reason, researchers make solutions or ways to speed up the process. In this context, create a program that can estimate how effective a drug is against the existing type of cancer. This paper focuses on this aspect but goes deeper in terms of the algorithm used to program the prediction of drug sensitivity to cancer cell lines. The dataset used to train the algorithm is the GDSC dataset, specifically the dataset on liver cancer or LIHC. There are three algorithms analyzed, namely, MLP, Random Forest, and LightGBM. The results obtained are that the LightGBM algorithm with Grid Search preprocessing has the lowest RMSE value and the highest R-Squared value compared to the other two algorithms. With these results, it concluded that LightGBM is an effective algorithm for predicting drug sensitivity to cancer cell lines. For further future research, it is expected to explore other models such as ensembles and find out their effectiveness.

*Keywords:*LIHC, MLP, Random Forest, LightGBM , Grid Search, K-Fold, MinMaxScaler

1. Introduction

High-throughput screening of large numbers of molecules is a widely used approach to identify lead compounds that exert beneficial effects on certain phenotypes. When put into the context of cancer, chemical entity libraries have been tested in this way against a panel of cell lines grown under different conditions and with heterogeneous genomic backgrounds [1]. This test can also be done because there is a dataset that can be tested. The cell lines in the CCLE and GDSC data sets are derived from several human cancer tissues, such as the lung, breast, and kidney. The CCLE and GDSC datasets have abundant genomic data that includes gene expression, DNA copy number, mutation oncomap, and so on. These high-throughput data sets and screening have made major contributions to cancer biology and cancer treatment research [2].

In this research, various algorithms are used to predict drug sensitivity to existing types of cancer. Predictions can be achieved through algorithm training with the dataset used. Compared to other studies, this paper focuses more on comparing the capabilities of each algorithm and knowing the advantages or disadvantages of each algorithm through the results. The results sought are RMSE and R-Squared are acquired from three algorithms, namely Multi-Layer Percepton Neural Network, Random Forest, and LightGBM. This research also focuses on a type of liver cancer, Liver hepatocellular carcinoma (LIHC). The data obtained through Genomics of Drug Sensitivity in Cancer (GDSC) is one of the popular datasets used for cancer studies[3].

2. Area and Data

2.1 Previous Works

1. Menden, M. P., Iorio, F., Garnett, M. J., McDermott, M. W., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLOS ONE, 8(4), e61318. https://doi.org/10.1371/journal.pone.0061318

   Prediction of cancer cell response to drugs is the main goal in modern oncology. Ultimately, this predictive drug/therapeutic response to cancer cells will lead to cancer treatment. In this study, they developed machine learning models to predict the response of cancer cell lines to drug treatment, measured through IC50 values. They use GDSC dataset of drugs and for each drug, a neural network model was trained to predict its IC50 profile across the panel of cell lines based on the genomic background of each cell. Models predicted IC50 values in a 8-fold cross-validation and an independent blind test with coefficient of determination R2. The range of predicted IC50 values for a drug are typically narrower than for the observed values and is likely because currently available genomic dataset are in sufficient to explain the observed range of drug responses across the cell lines.

2. Minteer, S. D., Wang, Y., Zheng, R., Shi, X., Li, Y., Wu, F., & Wang, J. (2021). DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 575–582. https://doi.org/10.1109/tcbb.2019.2919581

   Cancer lines with heterogeny genomic backgrounds are materials to study the molecular basis of drug activity and will lead to discover anticancer drugs. Institute such as National Cancer Institute(NCI), The Cancer Cell Line Encyclopedia (CCLE) and The Genomics of Drug Sensitivity in cancer (GDSC) have dataset from the effort to screen tumor cell line, compiled and catalogue genomics profiles. These datasets have made great contribution to cancer research and treatment. Computational approaches to analyzing drug sensitivity data can benefit to developing new anticancer drugs In this paper, they developed a deep learning architecture and improve the performance of drug sensitivity prediction base on the dataset. It is called DeepDSC and it will predict drug sensitivity of cancer cell lines by integrating genomic profiles of cell lines and chemical profiles of compounds. They use two datasets (CCLE and GDSC) and performed 10 fold-cross-validation to the deep learning models. They also compared DeepDSC with the state of the art method on CCLE dataset. DeepDSC outperformed state of the art method with the lowest prediction errors. DeepDSC also has hight coefficient determination.

3. Pouryahya, M., Oh, J. M., Mathews, J. M., Belkhatir, Z., Moosmüller, C., Deasy, J. O., & Tannenbaum, A. (2022). Pan-Cancer Prediction of Cell-Line Drug Sensitivity Using Network-Based Methods. International Journal of Molecular Sciences, 23(3), 1074. https://doi.org/10.3390/ijms23031074

   The study of drug sensitivity to disease has recently made significant advances driven by advances in technology that can produce large volumes of biological data at low cost. Pioneering datasets include the NCI-60 database, the Genomics of Drug Sensitivity in Cancer (GDSC) project, and the Cancer Cell Line Encyclopedia (CCLE) project. Development of a predictive model that can be relied upon to identify cancer drugs that are optimal for cancer cells is an important step to accelerate the success of cancer treatment. Therefore, in this research they propose a network-based methodology that breaks down the problem into smaller, interpretable problems to improve the predictive power of anticancer drug responses in cells. They used the GDSC database for 915 cells and 200 drugs, the 915 cells divided into six clusters. Random forest modeling was carried out for each pair of cell-line drug clusters. In comparison, they also performed a threefold schema cross-validation through a random forest on the entire dataset without prior clustering. In conclusion, predictive modeling of clusters that have homogeneous data tends to increase the accuracy of predicting drug sensitivity. The limitation of this study is that using IC50 as a measure of drug

sensitivity has the possibility of bias. This is due to differences in growth rates of growing cancer cells and changes in the number of control cells during the observation period.

4. E. Guo, P. Torabi, D. E. Nielsen, and M. Pietrosanu, "Deep learning transcriptomic model for prediction of pan-drug chemotherapeutic sensitivity," vol. 7, no. 1, pp. 40–53, Jan. 2022, doi: https://doi.org/10.17975/sfj-2021-013.

   Pan-cancer and pan-drug predictive biomarkers have good clinical utility in selecting therapeutic candidates for a particular patient. In the clinical context we can implement a Deep Learning approach to minimize the number of transcriptomic features in order to maximize the accuracy and interpretation of these predictive biomarkers. Biomarkers can also be useful for classifying patients into different treatment response groups thereby enabling chemotherapy regimens to be adapted to the transcriptomic characteristics of a particular cancer. In this study they used the GDSC method to classify types of cancer cells into therapeutic response groups. The GDSC database will later be used to understand the impact and predictive ability of transcriptomic dysregulation in chemotherapy response where the database includes 1110 cell lines from various types of tumors and pan-cancer data sets. This study also implements a dataset that contains information on therapeutic efficacy in the form of half-maximum inhibitory concentration values (IC50) for 251 types of chemotherapy. Then after going through several processes, it was concluded that the neural network with 10 hidden layers is the most accurate model with response prediction results on cell line therapy reaching an accuracy of 89.0%.

5. F. Xia *et al.*, "A cross-study analysis of drug response prediction in cancer cell lines," vol. 23, no. 1, Sep. 2021, doi: https://doi.org/10.1093/bib/bbab356.

   In providing an assessment for drug response prediction, this paper focuses on 5 cell line screening studies by integrating different drug response matrices and applying deep learning methods. In the application of data integration, this study applies the prediction of single drug response and uses 5 datasets such as: NCI60, CTRP, GDSC, CCLE, gCSI. Then the response prediction upper bounds show results that are biased or not fixed. In terms of response variability, this study selected 3 independent response metrics such as AUC, AAC, and DSS which would later be used to analyze the variability of cross-study responses by focusing on common subsets and cell lines of drug pairs that appeared in multiple studies. Implementation of the Deep Learning learning model is then carried out to evaluate the context of a single data set, then it is carried out based on 2 general factors such as input features and model complexity. Then, in the feature selection and processing method, the study uses 3 basic features, namely: RNAseq gene expression profile, drug descriptors and drug fingerprints. As for the application of machine learning, there are several methods or algorithms used, such as RandomForest which is trained using the default implementation of Scikit-learn, LightGBM which is trained using the implementation of Scikit-learn with the number of set trees proportional to the number of drugs included in the training set, and deep neural. networks. And after going through several processes, it is concluded that the variability of drug response within and across studies limits performance in machine learning in a biased way, and in the deep learning process it has promising results and in terms of maximizing the value of screening data, it is necessary to increase model generalization and standardization in drug diversity.

2.2 Dataset

The data of drug sensitivity to cancer cell lines are obtained from laboratory processes, which starts with the selection of cancer cells to be tested, then exposing them to a number of drugs, and assessing the condition of the cancer cells. The obtained data are then analyzed to determine the sensitivity of the drugs to cancer cell lines [2].

The dataset used in the experiment is obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) website, where the dataset consists of 806 rows of data. The dataset was downloaded with the specification of Liver Hepatocellular Carcinoma (LIHC) tissue. There are 274 different types of drugs tested in the dataset, with 25 different target pathways [3].

3. Methodology

3.1. Preprocessing

- Data Normalization - MinMaxScaler

  Normalization of data using MinMaxScaler is a process of transforming data to a specific range, typically between 0 and to ensure equal contribution from all features in the final result [4].

  The MinMaxScaler used in this research is obtained from the Sklearn library in the preprocessing package. This method of data normalization is based on the concept that a dataset with variables or features that have values in different ranges does not have equal weight or contribution in the final calculation [4]. By using MinMaxScaler, the limitations of various machine learning algorithms in handling different or wide-ranging variable values can be effectively addressed.

- K-Fold

  K-Fold is a machine learning technique used to evaluate the performance of a model. K-Fold works by dividing or splitting the dataset into K equal parts [7]. Each part is trained with a separate model. These models' performances are then averaged.

  The advantage of K-Fold is that it utilizes the entire dataset as both training and testing data, compared to splitting the data into separate training and testing sets in normal process without K-Fold [5].

  The K-Fold used in this experiment is Stratified type K-Fold obtained from the Sklearn library in the model selection package. Stratified K-Fold ensures that the distribution of classes in each fold is balanced [8].

- Hyperparameter Tuning - Grid Search

  Hyperparameter Tuning is an optimization method for machine learning models. This method searches for the best set of hyperparameters for the machine learning model [9]. Hyperparameters are parameters that cannot be searched using data, so they must be specified in the model before conducting model training. Some examples of these hyperparameters are the learning rate, the number of hidden layers in the neural network model, and others. The goal of Hyperparameter Tuning is to find the right balance so that the resulting model is neither overfitting nor underfitting.

  One of the most popular methods of hyperparameter tuning is the Grid Search. Grid Search is counted as a brute force method, in which the system will train the model with all possible hyperparameter combinations. The hyperparameter set that works best will be chosen to be the best set and will be used [9]. Grid Search is a method that is relatively simple but takes a lot of resources and time to do.

3.2. Algorithm

- Multi-Layer Perceptron Neural Network

  Multi-Layer Perceptron (MLP) is one of many Artificial Neural Network (ANN) models. It is a feedforward neural network model consisting of multiple interconnected layers of nodes, known as neurons. MLP is commonly used in tasks such as classification, regression, and pattern recognition.

  The structure of MLP consists of an Input layer, one or more Hidden layers, and an Output layer. Each layer consists of neurons or units, and each neuron in one layer is connected to neurons in the next

layer. Information flows through the network in a forward direction, from the Input layer to the Output layer [1].

Training an MLP involves two steps: forward propagation and backpropagation. In forward propagation, input data is passed through the network, and output data is computed. The computed output is then compared to the desired output, and the error is calculated. In backpropagation, the error from forward propagation is propagated back through the network, and the network optimizes the weights of connections between nodes and layers to minimize the error. This process is repeated until the model achieves satisfactory performance. MLP is a versatile model but is prone to overfitting if not properly regulated.

- Random Forest

Random Forest is a popular supervised machine learning algorithm used for classification and regression tasks [10]. It operates by creating an ensemble of decision trees that are individually diverse yet accurate. The key idea behind Random Forest is to train each decision tree on a random subset of the training data, reducing the risk of overfitting and promoting diversity among the trees.

In classification cases, Random Forest combines the predictions of the individual decision trees through majority voting, where the most frequent prediction among the trees is selected as the result. In regression cases, the predictions from each tree are averaged to obtain the result.

- LightGBM

LightGBM is a gradient boosting machine used for regression and classification tasks. The name LightGBM stands for Light Gradient Boosting Machine. While LightGBM and XGBoost share the concept of dividing data into subsets and building decision trees using boosting techniques, there are differences in their implementation [6].

LightGBM creates decision trees by adding new leaves, not new branches. This approach has the potential to lead to a significant reduction in error.

LightGBM is designed to be efficient and effective in handling complex data by optimizing memory usage and computational speed. This allows LightGBM to process larger datasets and save time compared to traditional gradient boosting methods.

In summary, LightGBM is a powerful algorithm that leverages gradient boosting techniques to build accurate and efficient models for regression and classification tasks.

4. **Result**

This experiment uses the dataset from the Genomics of Drug Sensitivity in Cancer (GDSC) with 22 features. The dataset is used in its entirety as both training and testing data using K-Fold. In addition to K-Fold, the preprocessing techniques used in this experiment are MinMaxScaler, and Hyperparameter Tuning.

Multi-Layer Perceptron Neural Network (MLP) model is used with 5 input nodes, 64 hidden nodes, and 1 output node.

Random Forest model is used with the parameter n_estimators set to 100 and max_depth set to 5.

LightGBM model is used with the parameter num_boost_round set to 100.

The results from all models are presented in the metric of Root Mean Squared Error (RMSE) and R-squared.

RMSE represents the average difference between the predicted values from a regression model and the actual observed values. RMSE is calculated by taking the square root of the average of the squared differences between the

predicted and observed values. RMSE indicates the accuracy of the model, with a lower value indicating better performance.

R-squared (R^2) represents the proportion of variance in the dependent variable that is explained by the independent variables in the regression model. It ranges between 0 and 1 and is typically expressed as a percentage (%). In contrast to RMSE, R-squared measures the accuracy of the model, with a higher value indicating better performance.

Experimental Result

	RMSE	R-squared
MLP	1.93 +/- 0.14	0.24 +/- 0.06
Random Forest	1.97 +/- 0.14	0.21 +/- 0.06
LightGBM	1.96 +/- 0.16	0.22 +/- 0.08

Table 1 : Result of Algorithms without Grid Search

	RMSE	R-squared
MLP	1.88 +/- 0.13	0.28 +/- 0.05
Random Forest	1.93 +/- 0.18	0.25 +/- 0.06
LightGBM	1.86 +/- 0.13	0.30 +/- 0.05

Table 2 : Result of Algorithms with Grid Search.

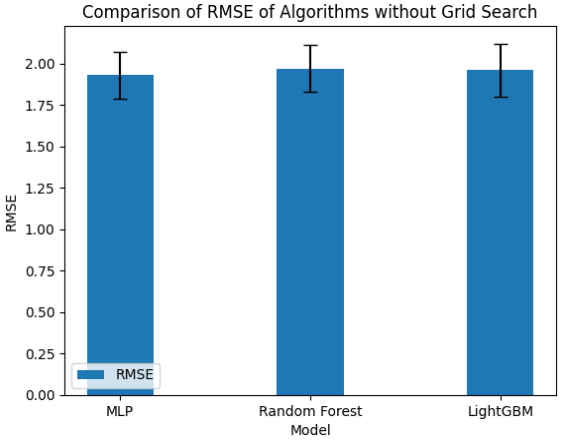

Fig 1.1 : RMSE values of Algorithms without Grid Search.

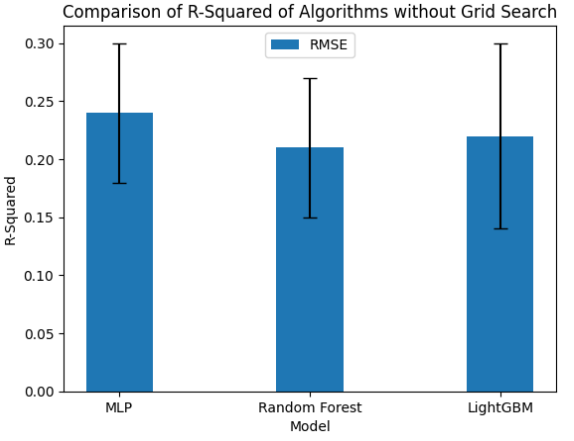

Fig 1.2 : R-Squared values of Algorithms without Grid Search.

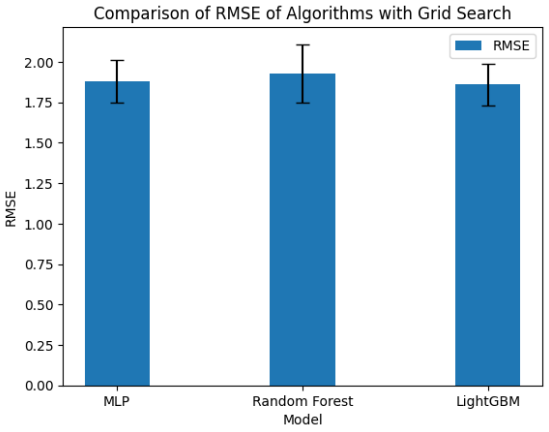

Fig 2.1 : RMSE values of Algorithms with Grid Search.

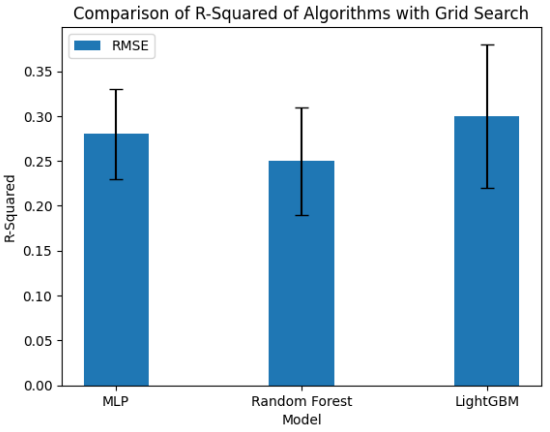

Fig 2.2 : R-Squared values of Algorithms with Grid Search.

From Table 1, it can be concluded that without the implementation of Grid Search, Multi-Layer Perceptron (MLP) yields the best results with RMSE of 1.93 +/- 0.14 and R-squared of 0.24 +/- 0.06. This is followed by LightGBM with RMSE of 1.96 +/- 0.16 and R-squared of 0.22 +/- 0.08. Finally, Random Forest achieves RMSE of 1.97 +/- 0.14 and R-squared of 0.21 +/- 0.06.

Therefore, it can be concluded that the MLP model performs the best compared to the others in a situation without Grid Search.

From Table 2, we can conclude that with the implementation of Grid Search, LightGBM produces the best results with RMSE of 1.86 +/- 0.13 and R-squared of 0.30 +/- 0.05. This is followed by MLP with RMSE of 1.88 +/- 0.13 and R-squared of 0.28 +/- 0.05. Finally, Random Forest achieves RMSE of 1.93 +/- 0.18 and R-squared of 0.25 +/- 0.06.

Therefore, it can be concluded that the LightGBM model performs the best compared to the others with the implementation of Grid Search.

From the tables above, we can conclude that LightGBM can produce the best results with the help of Grid Search. Then, followed by the MLP model, which is capable of achieving the best results without Grid Search. Random Forest has the least satisfactory results in all experiments compared to the other two algorithms but still produces good results. Additionally, Grid Search plays a role in improving the results of all tested algorithms.

5. Conclusion

In this experiment, our aim was to predict drug sensitivity to cancer cell lines using the GDSC dataset and the LightGBM algorithm. The results of LightGBM experiment shows that with the implementation of preprocessing, in this case, Grid Search, LightGBM is able to achieve the lowest RMSE value and the highest R-squared value compared to Multi-Layer Perceptron (MLP) and Random Forest. In addition to evaluating multiple algorithms, we also implemented Hyperparameter Tuning through Grid Search, which reduced the RMSE value by 0.1 and increased the R-squared value by 0.08.

The best results were obtained from LightGBM experiment with Grid Search, with RMSE value of 1.86 +/- 0.13 and R-squared value of 0.30 +/- 0.05.

The results of this experiment demonstrate that LightGBM can effectively predict drug sensitivity to cancer cells using the GDSC dataset and outperform the other two algorithms. Our suggestion for future research is to explore the effectiveness of ensemble models such as stacking and boosting methods.

Acknowledgements

We thank the School of Computer Science of Bina Nusantara University, especially Genomic of Drug Sensitivity in Cancer Cell for providing research ideas and computational resources; Bina Nusantara University Project Initiative (PID36192): Nationwide Integrated Conference Ecosystem that has funded this research.

Reference

[1] Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS ONE, 8(4), e61318. https://doi.org/10.1371/journal.pone.0061318.

[2] Li, M., Wang, Y., Zheng, R., Shi, X., Li, Y., Wu, F., & Wang, J. (2019). DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1. https://doi.org/10.1109/tcbb.2019.2919581.

[3] Wellcome Sanger Institute & Massachusetts General Hospital Cancer Center. (2023). Genomics of Drug Sensitivity in Cancer Project. Retrieved from https://www.cancerrxgene.org.

[4] E. Guo, P. Torabi, D. E. Nielsen, and M. Pietrosanu, "Deep learning transcriptomic model for prediction of pan-drug chemotherapeutic sensitivity," vol. 7, no. 1, pp. 40–53, Jan. 2022, doi: https://doi.org/10.17975/sfj-2021-013.

[5] F. Xia *et al.*, "A cross-study analysis of drug response prediction in cancer cell lines," vol. 23, no. 1, Sep. 2021, doi: https://doi.org/10.1093/bib/bbab356.

[6] Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). doi:10.1109/icssit48917.2020.9214160.

[7] Nematzadeh, Z., Ibrahim, R., & Selamat, A. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In 2015 10th Asian Control Conference (ASCC). https://doi.org/10.1109/ascc.2015.7244654.

[8] Wang, D., Zhang, Y., & Zhao, Y. (2017). LightGBM. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics - ICCBB 2017. doi:10.1145/3155077.3155079.

[9] Pal, K., & Patel, B. V. (2020). Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). https://doi.org/10.1109/iccmc48092.2020.iccmc-00016.

[10] Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology, 4, 972421.

[11] Shekar, B. H., & Dagnew, G. (2019). Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. In 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP) (pp. 1-6). doi:10.1109/icaccp.2019.8882943.

[12] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. Expert Systems with Applications, 131, 213-222. doi:10.1016/j.eswa.2019.05.028.

[13] Menden, M. P., Iorio, F., Garnett, M. J., McDermott, M. W., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLOS ONE*, *8*(4), e61318. https://doi.org/10.1371/journal.pone.0061318

[14] Minteer, S. D., Wang, Y., Zheng, R., Shi, X., Li, Y., Wu, F., & Wang, J. (2021). DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 575–582. https://doi.org/10.1109/tcbb.2019.2919581

[15] Pouryahya, M., Oh, J. M., Mathews, J. M., Belkhatir, Z., Moosmüller, C., Deasy, J. O., & Tannenbaum, A. (2022). Pan-Cancer Prediction of Cell-Line Drug Sensitivity Using Network-Based Methods. *International Journal of Molecular Sciences*, *23*(3), 1074. https://doi.org/10.3390/ijms23031074

[16] E. Guo, P. Torabi, D. E. Nielsen, and M. Pietrosanu, "Deep learning transcriptomic model for prediction of pan-drug chemotherapeutic sensitivity," vol. 7, no. 1, pp. 40–53, Jan. 2022, doi: https://doi.org/10.17975/sfj-2021-013.

[17] F. Xia *et al.*, "A cross-study analysis of drug response prediction in cancer cell lines," vol. 23, no. 1, Sep. 2021, doi: https://doi.org/10.1093/bib/bbab356.