Leveraging Machine Learning to Identify Synergistic Drug Combinations for Effective Cancer Treatment

P. Sujatha,
*Research Scholar,
Department of AI/ML,
Amrita Vishwa Vidyapeetham,
Amaravati, Guntur,India,*
P_sujatha@av.students.amrita.edu

K Saravanan,
Associate Professor, Department of Information Technology, R.M.K Engineering College, Tiruvallur District,Tamil Nadu,India,
ksn.it@rmkec.ac.in

Mohammed Ali Sohail,
Lecturer, Department of Computer & Network Engineering, College of Computer Science & Information Technology , Jazan University, Jazan, K.S.A,
msohail@jazanu.edu.sa

Basi Reddy.A,
Assistant Professor, Department of Computer Science and Engineering, School of Computing, Mohan Babu University, Tirupati,Andhra Pradesh,India,
basireddy.a@gmail.com

Rohit R Dixit,
*Associate Professor,
Senior Data Scientist, Siemens Healthineers, New Hampshire-03755, USA,*
RohitDixit188@live.in

Nallam Krishnaiah,
Professor, Department of AI&ML, St.Martin's Engineering College Dhulapally, Secunderabad Telangana 500100,India,
nkrishna520@gmail.com

2023 Seventh International Conference on Image Information Processing (ICIIP) | 979-8-3503-7140-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIIP61524.2023.10537707

Abstract-**The potential of drug combinations to treat and overcome medication resistance complex genetic diseases is evident. Synergistic drug combinations offer a promising way to enhance drug therapy efficacy and reduce the required medication dosage. However, developing effective combination medication therapies with synergistic effects has been challenging, despite numerous ongoing clinical investigations. Current models and approaches to detect medication synergy outlined in the literature lack the expected consistency in outcomes. to better comprehend the impact of particular medication combinations, it is essential to be familiar with the vocabulary used to describe synergy. In this study, a combinational drug screen is utilized to identify useful features for locating synergistic or efficient drug combinations. The feature selection algorithm (Boruta) helps select the most relevant features, and machine learning models are then trained using the selected feature dataset. Performance assessment metrics like sensitivity, accuracy, and specificity are used to compare the trained models, and the Random Forest model stands out for its significantly better performance compared to other models.**

Keywords: Drug combination, Drug resistance, Synergy, Efficacy, Machine learning and Random Forest

I. Introduction

Cancer has become a leading global cause of death, and anti-cancer drugs play a crucial role in treatment, extending patients' lifespans. However, despite identical therapies, due to genetic differences, patients with an identical cancer type frequently respond differently[1]. Gene-specific targeted therapy has been proposed as a potential solution for cancer treatment. Still, it requires extensive research studies, which face challenges such as limited samples, complex procedures, strict environmental requirements, and high costs[2]. To address these issues, to develop models that can forecast the effects of drugs, researchers have combined genomic data and data on drug response. For instance [3] developed a regression model using the random forest method, successfully predicting drug responses in breast cancer and glioblastoma cell lines.Other techniques, like those based on asymmetrical gene expression, have also been explored. [4] to predict clinical medication response, baseline expression levels of genes and in vitro drug sensitivity are taken into account. Interestingly, studies have revealed that structurally [5] similar pharmaceuticals can produce similar effects on cancer cell lines with comparable gene expression profiles. Building on this concept, a new enhanced system for predicting drug reactions based on cancer genomics has been developed, identifying potential predictor genes for drug response through data analysis.

II. Methods and Materials

This research made use of data from the Genetics of Drug Sensitivity in Cancer (GDSC) database[6], created by the Sanger Institute in the United Kingdom. The study focused on 12 different medications and gene expression data collected from one thousand human cancer cell lines[7]. Two key indicators, an inhibitory level that is half maximum (IC50) and the area beneath the curve (AUC), were employed to evaluate drug response. The AUC stands for the area that lies beneath the dose-response curve, whereas the IC50 specifies the medication level which reduces cell viability by 50% [8]. Lower IC50 and AUC values correspond to a higher tumor cells' reaction to the medication. To predict drug responses, the researchers used a machine learning algorithm and the pair of response indicators [9] based on cancer cell line gene expression data. Initially, they employed Elastic Net to make predictions using genes selected through the p-value of the Pearson correlation coefficient. Additionally, separate Elastic Net regressions were performed for each response value, and predictor genes were chosen from the previously identified genes[10] . The aim was to discover common predictor genes capable of predicting drug response with higher or comparable accuracy to the independently derived results for both response indicators. To determine the biological relevance of the predictive genes[11], map and gene taxonomy analyses were conducted. Fig.1. illustrates the complete experimental procedure.

A. *Features Selection Based on Pearson Correlation Coefficient*

In the gene expression data of certain pharmaceuticals, there are numerous genomes, but only a few genes show a strong correlation with drug responses. To ensure the selection of relevant genes, a pre-selection step becomes crucial. However, Elastic Net, which is capable of gene selection, can be influenced by data dependencies or bulk effects, leading to potential errors [12]. This might result in the exclusion of genes that are vital in predicting drug responses.To overcome this issue, a two-step gene selection approach was implemented. Initially, genes were preselected using the Pearson correlation coefficient before applying Elastic Net. This allowed for a more refined selection of predictor genes by considering the p-value of the connection between medication reaction and gene expression [13-14]. Genes with a significance level of 0.05 or lower were chosen during the initial feature selection process. Elastic Net is a l1 and l2 regularised linear regression framework that is particularly helpful when working with a large number of linked features. Given that our dataset has significantly more features (genes) than samples, there is a risk of overfitting in the prediction. To tackle this, Elastic Net was employed to select genes and reduce the generalization error when predicting drug response.

B. Elastic Net-Based Feature Selection and Drug Response Prediction

We conducted exploratory experiments to assess Elastic Net's suitability by comparing it with two well-known approaches [15]. SVR can function as a non-linear regressor with different our experiments, we used the radial basis function kernel as a kernel function. Xgboost, on the other hand, is an enhanced version of the gradient boosting algorithm based on decision trees. While both algorithms have shown good performance in various applications, our preliminary experiments indicated that they were more prone to overfitting compared to Elastic Net. Elastic Net fared better than SVR and Xgboost when comparing the quantity of shared predictive genes for the two response indicators. Using Pearson relationship coefficients to contrast the anticipated results with and measured IC50 values, Fig. 2. presents a summary of the comparison between the twelve medicines. The figure presents a comprehensive outline of the proposed framework for predicting drug synergy and efficacy. The process begins by collecting data from single-agent and multiple-agent substance screenings. Specifically, Using the procedure outlined in section 5.2.1, IC50 values are gathered on 27 cell lines from single-agent drug testing. These IC50 values are then normalized within the [-1, 1] range with log normalisation, where -1 represents the least sensitive drug-cell-line pair, and 1 indicates the most resistant pair. Features to analyse the synergy and efficiency of combination drug therapy are established based on data on single-agent drug response. The desired data undergoes normalization using the min-max normalization technique, which ensures a linear transformation of the input information while preserving the original relationships between data values. Next, the feature selection algorithm, Boruta, is employed to rank various features based on their average reduction in the Gini coefficient. Features with the highest and average Gini importance are selected. The prepared dataset is used to train a variety of machine learning models in the fourth stage. The optimal model is then selected by comparing performance indicators such as sensitivity, specificity, accuracy, and area under the curve (AUC). Once the finest model for predicting drug synergy and efficacy is chosen, model testing is performed. Combinations with a combination index (CI) of -1, resulting in a growth suppression rate of 70%, are considered efficacious and synergistic.

C. Phase-1: Features Selection using Boruta

For further validation, two strategies are employed. To begin with, we do a 10-fold cross validation and use effectiveness parameters like sensitivity and specificity to evaluate the model's resilience. The data is split into 10 equal groups for this operation, using the other groups as the training set and one group as the test set. To lessen bias between various data occurrences during training and testing, the testing findings are then averaged. In the (1) estimate efficient and synergistic drug combinations, a second dataset is utilized for additional validation purposes.

Imp(X_)=1/N_TΣ_ 〖p(x)〗
Σ_(X∈T:V(S_X)=X_) p(x)Δi(s_x,x)〗 (1)

The proportion p(x) denotes the ratio of Nt/N samples that visited vertex x from vertex v(sx) and were used for dividing Sx. When the Gini index is used as the measure of impurity, it is referred to as the mean decrease Gini coefficient. Based on their highest and average Gini importance, a total of 24 features were selected. Fig. 3. visualizes the mean decrease in the Gini coefficient for these chosen features, where a lower Gini value indicates higher significance.

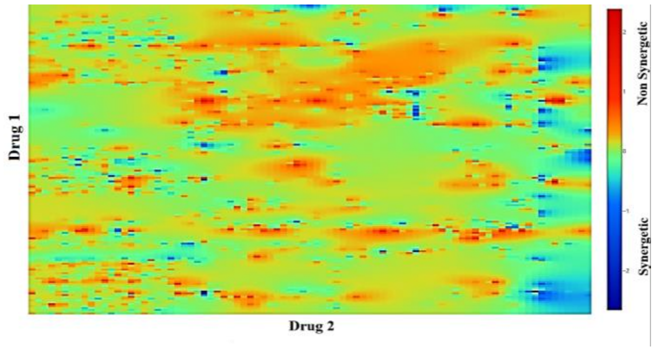

Fig. 1. Heat map: Drug combination of DREAM challenge dataset

Table 1 presents the nine different machine learning models used in this research, all of which are available as R packages, a programming language that is open-source and covered by the GNU General Public License. The table also provides details of the tuning parameters utilized during model training.The initial step in the analysis involves IC50 normalization for drug response values. Resistant cell lines have responses (IC50) higher than the maximum drug concentration, while sensitive cell lines have lower responses. To achieve IC50 normalization, drug responses are divided by the highest concentration and then subjected to log2 transformation. This process results in obtaining a summary of all responses within the [-1, 1] range for the future analysis. The feature selection method, Boruta is used to rank the features in accordance with various characteristics determined by the mean decrease Gini coefficient. From this analysis, a selection of 24 features is obtained with the highest and average Gini importance.The chosen set of features is then used to train the machine learning models listed in Table 1. Among these models, the optimal one (random forest) is selected by comparing area under the curve (AUC), sensitivity, specificity, and accuracy are examples of performance indicators.

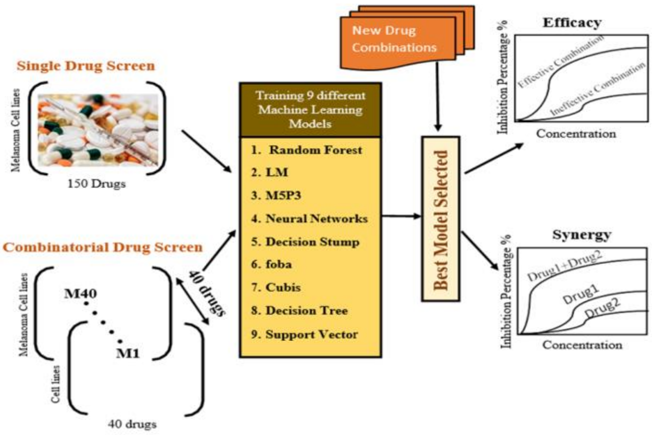

Fig. 2. Overall flow of the proposed work

Table 1. Parameters of the ML models

ML Model	Equipment Used
(Random Forest)	mtry=2.0, ntree=500.0, sampling=bagging
(Neural Network)	size=10.0
(Support Vector)	epsilon=0.6,nu=0.4
LM	None
M5P3	pruned=15.0, rules=9.0,smoothed=0.6
(Decision Stump)	smoothed=0.9,rules=6.0,pruned=25.0
Foba	lambda=1e-5, k=1000
Cubis	neighbors=30.0,ommittees=10.0
(Decision Tree)	maxdepth=30.0 ,minsplit=20.0,minbucket=7.7

III. Result and Analysis

A. Drug-Drug Similar Matrices

In this study, the predictive model's durability is assessed using K-fold cross-validation, which divides the dataset into K pieces, and (K-1) groups are used for training the model while one subset is used for evaluation. This process is repeated for each subsample, generating K results, and the final output is obtained by averaging these outcomes. A 10-fold cross-validation is employed in this research, dividing the dataset into ten equal-sized parts for validation, ensuring robustness. The primary goal of the research is to predict the effectiveness and synergy of drug combinations based on the dose response of a single agent. Previous studies have shown that the dose-response curves of individual agents offer insights into combinational responses. This is accomplished by applying the Held et al. high-throughput drug screen, which focuses on 150 dosage responses data points for a single agent with 40 compounds in the context of BRAF-melanomas. 27 cell lines are used in the drug screening, including cell lines with RAS, RAS(WT), mutant BRAF, and BRAF(WT) mutations. Additional details on some of the pharmaceuticals used in Held et al.'s drug combination screening can be found in Table 2. Based on the dose-response of single agents, the median and range of each response parameter for each cell line are calculated for each drug combination. Feature extraction is vital as it plays a crucial role in computational models used to predict drug interactions. Identifying highly responsive features enhances prediction accuracy and provides insights into the underlying mechanisms of synergy. In this study, features are calculated to leverage the genetic effects of each medicine[1] on the number of responses for each cell line combinations.

Table 2. Parameters of the ML models

Model Name	Accuracy	Specificity	Senstivity
(Random Forest)	0.8212	0.9081	0.7942
(Neural Network)	0.7233	0.7104	0.8572
(Support Vector)	0.7768	0.8899	0.7501
LM	0.7546	0.7354	0.8183
M5P3	0.7112	0.7134	0.6971
(Decision Stump)	0.6567	0.6783	0.5653
foba	0.7012	0.7151	0.6251
cubis	0.6810	0.6892	0.6251
(Decision Tree)	0.6979	0.7501	0.6890

In order to indicate the similarity of pharmacological combinations for each combination, we have identified 54 features. As shown in Table 1, we first used a dataset made up of 750 perturbed RAS and BRAF melanoma medication combinations to train several machine learning models. Based on performance metrics including accuracy, specificity, and sensitivity, Table 2 compares these models. To reduce bias caused by the training-testing partition, each model is trained using four distinct partition sets, emphasizing the consistency of the models across diverse partition sets. In both tables, the first position in each entry represents accuracy, while the second position represents sensitivity. The results clearly indicate that the random forest model outperforms other methods. The random forest model emerges as the most effective training model for machine learning, exhibiting higher accuracy and specificity, indicating a lower error rate or incorrect predictions. Specifically, according to reports, the random forest synergy model's accuracy and specificity are 0.9091 and 0.8222, respectively. This demonstrates the superior performance of the random forest model in predicting drug synergy and efficacy in this study.

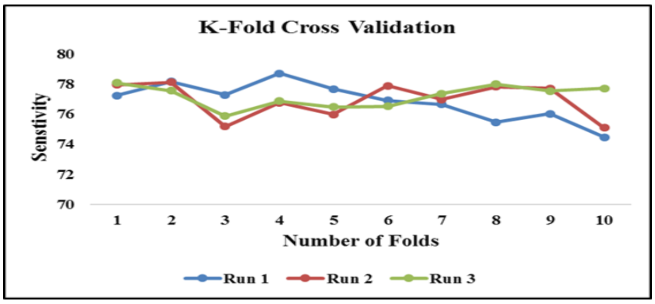

Fig. 3. Random Forest with K-fold (K=10) and cross validation Sensitivity

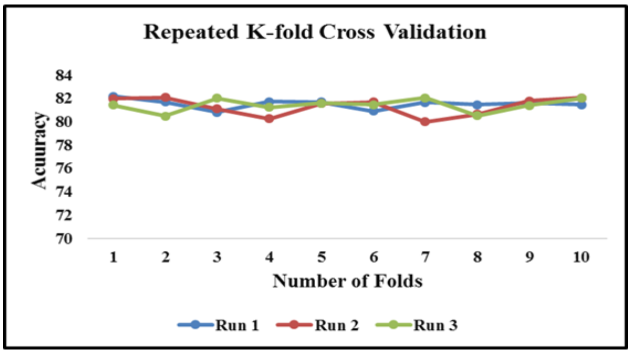

Fig. 4. . Using Accuracy and K-fold (K=10) cross validation for Random Forest

Next, we proceeded to train random forest models using the data outlined in section 5.2.1. Specifically, two separate models based on random forests were developed to predict the efficacy and synergy of drugs. The screening of combination drugs played a significant role in identifying drug combinations that exhibit synergy and are genotype-selective. Genotype-selective combinations refer to those that, on average, inhibit growth by at least 15%, with a typical inhibition level of at least 50%. Additionally, we defined an overall effective combination as one that inhibits growth by at least 70%. Fig. 3. and 4. presents the predictive performance of both models: the synergistic model (with an accuracy of 0.8222 and specificity of 0.9091) and the genotype-selective-effective model (with an accuracy of 0.8319 and specificity of 0.8963).

IV. Conclusion

In order to uncover synergistic or efficient drug combinations, this research uses a mixed drug screen to extract useful data. The study compares and assesses the effectiveness of machine learning models using the algorithm for selecting features (Boruta) and different machine learning models based on their sensitivity, accuracy, and specificity. Notably, when compared to other models, the Random Forest models have performed noticeably better. According to Table 1, 2, the accuracy of the Braf-synergy and Braf-effective models utilising the Random Forest model is stated to be 0.8319 and 0.8222, respectively. These predictive models align well with existing literature, indicating their potential in identifying effective medication combinations that work well together to treat certain malignancies. The proposed methodology demonstrates the potential to reduce the search space and effectively predict new drug combinations, leading to improved options for cancer treatment. The study's conclusions highlight the promising use of machine learning in drug combination research and its potential impact on advancing cancer therapies. By leveraging machine learning techniques, this research contributes to a better understanding of drug synergism mechanisms, opening doors to optimized and personalized treatment options for cancer patients. Moving forward, it is crucial to explore potential drug combination features to gain a holistic understanding of drug-disease interactions, ultimately enhancing the efficacy of cancer treatment. Additionally, the utilization of ensemble machine learning methods could further improve prediction performance in this context. Thereby, this research represents a significant step towards developing more effective and personalized cancer treatments through the identification of synergistic drug combinations. By utilizing machine learning techniques and incorporating genomic variability, this study contributes to the ongoing efforts to overcome drug resistance and improve cancer treatment outcomes, offering hope for more effective therapies and better patient outcomes in the fight against cancer.

References

[1] Pathania S., Bhatia R., Baldi A., Singh R., Rawal R.K. Drug Metabolizing Enzymes and Their Inhibitors' Role in Cancer Resistance. Biomed. Pharmacother. 2018;105:53–65. doi: 10.1016/j.biopha.2018.05.117

[2] Bijaideep Dutta, Sandeep B. Shelar, Ananya Nirmalraj, Sonali Gupta, Kanhu C. Barick, Jagriti Gupta, Puthusserickal A. Hassan. Smart Magnetic Nanocarriers for Codelivery of Nitric Oxide and Doxorubicin for Enhanced Apoptosis in Cancer Cells. ACS Omega 2023, 8 (47) , 44545-44557. https://doi.org/10.1021/acsomega.3c03734

[3] Yanglan Gan, Xingyu Huang, Wenjing Guo, Cairong Yan, Guobing Zou, Predicting synergistic anticancer drug combination based on low-rank global attention mechanism and bilinear predictor, Bioinformatics, Volume 39, Issue 10, October 2023, btad607, https://doi.org/10.1093/bioinformatics/btad607

[4] Y. Peng et al., "Electroencephalographic Network Topologies Predict Antidepressant Responses in Patients With Major Depressive Disorder," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 2577-2588, 2022, doi: 10.1109/TNSRE.2022.3203073.

[5] C. G. Manasseh, R. Veliche, J. Bennett and H. S. Clouse, "Static Seeding and Clustering of LSTM Embeddings to Learn From Loosely Time-Decoupled Events," in IEEE Access, vol. 11, pp. 64219-64227, 2023, doi: 10.1109/ACCESS.2023.3288487.

[6] W. Peng, T. Chen and W. Dai, "Predicting Drug Response Based on Multi-Omics Fusion and Graph Convolution," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 3, pp. 1384-1393, March 2022, doi: 10.1109/JBHI.2021.3102186.

[7] Onishi, I.; Yamamoto, K.; Kinowaki, Y.; Kitagawa, M.; Kurata, M. To Discover the Efficient and Novel Drug Targets in Human Cancers Using CRISPR/Cas Screening and Databases. Int. J. Mol. Sci. 2021, 22, 12322. https://doi.org/10.3390/ijms222212322

[8] Yuting Tang, Ting Wang, Yaowen Hu, Hongli Ji, Botao Yan, Xiarong Hu, Yunli Zeng, Yifan Hao, Weisong Xue, Zexin Chen, Jianqiang Lan, Yanan Wang, Haijun Deng, Chuxia Deng, Xiufeng Wu, Jun Yan, Cutoff value of IC50 for drug sensitivity in patient-derived tumor organoids in colorectal cancer, iScience,Volume 26, Issue 7, 2023, 107116, ISSN 2589-0042, https://doi.org/10.1016/j.isci.2023.107116.

[9] Ali, M., Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev 11, 31–39 (2019). https://doi.org/10.1007/s12551-018-0446-z.

[10] Aiedh Mrisi Alharthi, Muhammad Hisyam Lee, Zakariya Yahya Algamal, Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty, Informatics in Medicine Unlocked, Volume 24, 2021,100622, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100622.

[11] Q. Li and T. Milenković, "Supervised Prediction of Aging-Related Genes From a Context-Specific Protein Interaction Subnetwork," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 4, pp. 2484-2498, 1 July-Aug. 2022, doi: 10.1109/TCBB.2021.3076961.

[12] S. Song, F. Gao, R. Huang and C. Wang, "Data Dependencies Extended for Variety and Veracity: A Family Tree," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, pp. 4717-4736, 1 Oct. 2022, doi: 10.1109/TKDE.2020.3046443.

[13] Vohra, M., Sharma, A.R., Mallya, S. et al. Implications of genetic variations, differential gene expression, and allele-specific expression on metformin response in drug-naïve type 2 diabetes. J Endocrinol Invest 46, 1205–1218 (2023). https://doi.org/10.1007/s40618-022-01989-y

[14] Gonzalez, R.D.; Small, G.W.; Green, A.J.; Akhtari, F.S.; Motsinger-Reif, A.A.; Quintanilha, J.C.F.; Havener, T.M.; Reif, D.M.; McLeod, H.L.; Wiltshire, T. MKX-AS1 Gene Expression Associated with Variation in Drug Response to Oxaliplatin and Clinical Outcomes in Colorectal Cancer Patients. Pharmaceuticals 2023, 16, 757. https://doi.org/10.3390/ph16050757

[15] Román-Figueroa, C.; Cortez, D.; Paneque, M. A Comparison of Two Methodological Approaches for Determining Castor Bean Suitability in Chile. Agronomy 2020, 10, 1259. https://doi.org/10.3390/agronomy10091259