**Characterizing drugs for Pancreatic Cancer treatment using Machine learning models**

Gopikrishna Dasari, Karthik Reddy Marapureddy, Gowthami Dasari

Supervised by Othman Soufan

**Abstract**

Cancer is a chronic disease that can affect anyone. One in two Canadians is predicted to get cancer, and one in four Canadians is estimated to pass away from the disease each year, making it the leading cause of death in the country. A potential approach to treat complex diseases including cancer and infections is drug combination therapy. However, the current understanding of drug combination therapy, particularly for cancer patients, is insufficient due to toxicity, cell line heterogeneity, and unfavorable medication side effects. Despite an increase in cancer patients’ chances of survival, recurrence remains one of the disease's biggest challenges. Most cancer types are likely to return, even though recurrence rates vary greatly depending on the type of cancer, its stage, genetic factors, and the treatment received. Combination drug therapy, which employs two or more drugs that each function through a separate mechanism to help reduce the risk of cancer cells becoming resistant, is one method for treating recurring cancer[1]. Bioinformatics has evolved because of advances in biology and computer science. The difficulty of finding new medicine combinations to treat cancer was optimized using machine learning in this field. Building computational approaches, in particular machine learning techniques, could thus provide a successful plan to lower drug resistance and enhance treatment efficacy. In this research, we used machine learning models to identify potential drug combinations that may be useful in the treatment of pancreatic cancer.

**Key words:** potential; machine learning; drug combination therapy; bioinformatics.

**Introduction**

According to studies, one in four Canadians is affected by cancer, making it the country's biggest cause of death. About 65% of cancer patients have survived the disease in the past five years. Despite higher survival rates, more cases are being diagnosed each year because of the rising population. The 12th most common cancer in the world is pancreatic cancer. It ranks 11th among cancers in women and 12th among cancers in men. In 2020, there were around 495,000 new cases of pancreatic cancer. About 3% of cancers are caused by the disease. The eighth most frequent cancer in women and the tenth most frequent in men is pancreatic cancer. Since 2000, the annual increase in pancreatic cancer rates has been about 1%. In 2020, 495,773 cases of pancreatic cancer were expected worldwide. There are numerous therapies for pancreatic cancer, depending on how far along the illness is. There are several potential treatments, including surgery, chemotherapy, radiation therapy, and pharmacological combinations. Pancreatic cancer is largely treatable if discovered in its early stages. This is due to the fact that often times symptoms don't show up until the disease has spread to other organs. The cells that line the pancreatic ducts are where the majority of pancreatic cancers start. Pancreatic adenocarcinoma or pancreatic exocrine cancer are two terms used to describe this form of the disease. Less frequently, hormone-producing cells can develop cancer. These cancers are referred to as pancreatic neuroendocrine tumors or pancreatic neuroendocrine cells.

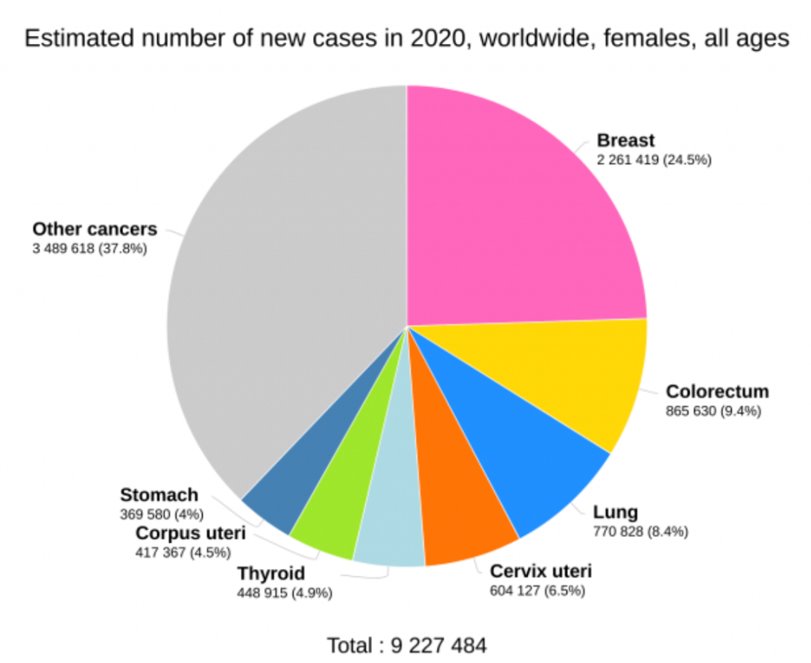


Fig 1

One of the big concerns is cancer recurrence, The person's cancer relapsing is one of the factors contributing to the rise in cases. Recurrence or recurrent cancer are phrases used in medicine to describe the disease's return in a patient. Some cancer cells become resistant to the treatment while being treated for their cancer. The time it takes for these resistant cancer cells to proliferate can range from a few weeks to several months or even years. One of the biggest challenges to treating cancer is the condition of recurrence. One of the big concerns is cancer recurrence which occurs when cancer returns after a while and stays undiagnosed. When certain cancer cells develop drug resistance and survive the treatment, this occurs.

Cancer can be treated in a variety of ways, but one of the most successful ways is using drugs. Additionally, the types of pharmacological therapy include combination drug therapy, personalized monotherapy, and one-size-fits-all monotherapy. When a patient has a comparable type of cancer, they will typically receive a single pharmacological treatment in a one-size-fits-all sort of treatment. According to the patient's cancer mutations, individualized monotherapy involves prescribing a specific drug. Patients are also undergoing drug combinations in addition to these treatments. Patients who receive this sort of treatment are given two or more different types of medications with various mechanisms, either in succession or simultaneously.

Various computational methods have been used to model this type of problem. Although there has been significant progress in the identification of drug combinations, no single computational approach is likely to be able to give all the necessary information in precise detail.

Every time a new medicine is released, the likelihood of discovering a new effective drug combination that would specifically target a characteristic of cancer increases. The field of medicine can benefit from the application of machine learning to predict potential drug combinations because there are enormous data sets from in vivo and in vitro clinical trials available. Various machine learning techniques have been employed in bioinformatics over time to classify data, train models, analyze large data sets, and predict synergistic drug combinations. Some of the significant methods are Knn, which is a greedy way to predict drug combinations, XGBoost, a method based on gradient-boosted decision trees, Random Forest, which can run effectively and handle large data sets, and Decision Tree, an easier and simpler approach, are some of the important methods. In this study, we predicted synergistic drug combinations for the treatment of pancreatic cancer using the above machine-learning models.

SMILES data, which is used to characterize the chemical structures of both the individual drugs and drug combinations, is another set of data that we have utilized. We have taken the data from [13] and extracted Lipinski's Rule of Five (Ro5) features, which we then integrated with the main dataset. The final dataset is created by extracting and combining a total of 5 features, including 1 experimental feature from the Drug Comb dataset and 4 features from the SMILES dataset.

The dataset is then partitioned into test and training data so that we can use different models in machine learning like XGBoost, Random Forest, Decision Tree Classifiers, knn are trained and used to predict synergistic drug combinations. The effectiveness of the 4 models is then compared, and the drug synergism values are discussed.

**Materials**:

We created a combined dataset by extracting features from the open-access data portal Drug Comb Portal. The dataset retrieved through Drug Comb is a raw dataset made up of 739,964 unique drug combinations, 2320 cell lines, 33 different types of tissues, 8397 different drugs, and 8397 different drugs evaluated on the same combinations across various in vitro studies. The dataset includes various features, including IC50 (half maximal inhibitory concentration), CSS (Combination Sensitivity Score), RI (Relative Inhibition), and CSS (Combination Sensitivity Score). IC50 is "the concentration of drug required for 50% inhibition; CSS is calculated on the combined dose-response curve's normalized area under the log10 transformed curve when one of the two. By extracting all of the canonical smiles for each and every drug in the dataset from the PubChem portal, we were able to create our own smiles dataset. When two or more drugs having similar effects are combined, the efficiency of those drugs increases. If the combined effect is working better, then their individual potencies would indicate synergy. Various models have different definitions for synergy leading to different synergy values for the same drug and cell line combinations.

**Method:**

**A) Dataset Preparation:**

The data from the Drug Comb portal is our base dataset. We filtered out only the values that were required for our use because the dataset contains values for a variety of cancer types. According to our requirements, we filtered the data and chose the NCAT study data. From the SMILES structure, we first extracted useful features. For each particular drug compound, we extracted 5 different types of features using RDKit. The basis for all five features is Lipinski's Rule of Five Descriptors.. The smiles dataset appears as follows:

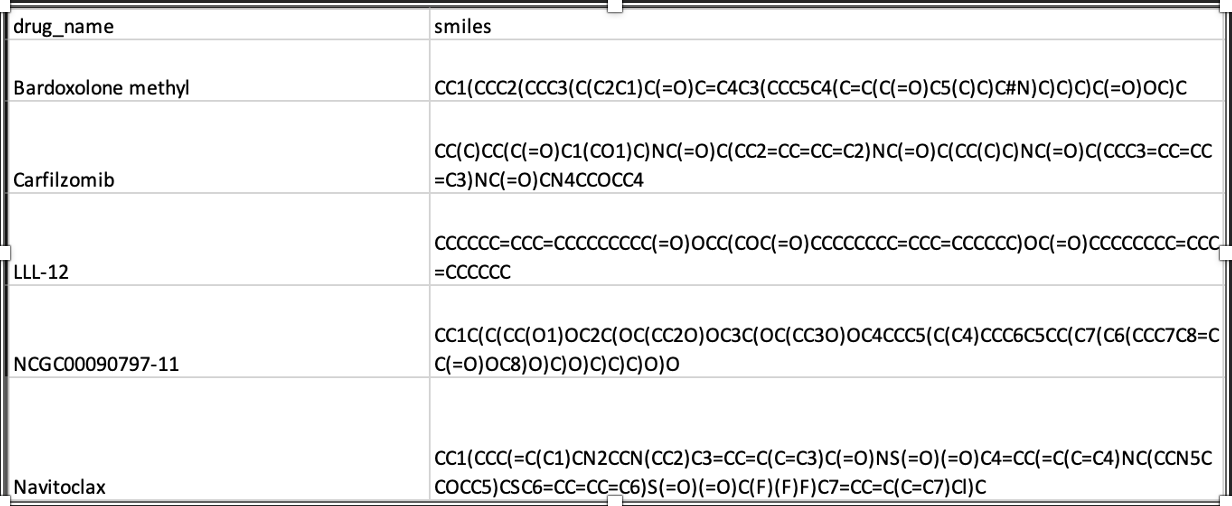


Table 1

The smiles data with Lipinski’s descriptors are given below:

Table

Description automatically generated

Table 2

Using Lipinski's rule of five descriptors, we determined the characteristics associated with drugs. Lipinski's rule describes a compound or molecule's likelihood of being a drug. The implementation of this approach enables the assessment of the probability that a biologically active molecule will have the chemical and physical characteristics necessary for oral accessibility. [21] The Lipinski rule focuses on certain molecular properties to determine drug pharmacokinetic parameters such as absorption, distribution, metabolism, and excretion such as

* 5 or fewer H (Hydrogen) bond donors
* 10 or fewer H bond acceptors
* A Mol Weight of 500 Da or less
* There must be no more than a five-partition coefficient.

When two or more of these characteristics are violated, a molecule is likely to be a non-drug that can be taken orally. Because all the above-mentioned conditions are multiples of 5, this phenomenon is known as the "rule of five." We are creating a dataset with all drug combinations and their molecular features, such as fingerprints, after first establishing a matrix with all drug combinations. The RDFingerprint function extracts fingerprints for each drug, and we use that data to calculate "DataStructs.FingerprintSimilarity" scores for each combination. Using the smiling data, we created a final dataset that we are merging with the primary dataset. The merged dataset is shown below.

Graphical user interface, table

Description automatically generated

Fig 2

Here,

* Drug\_row/Drug\_col: It refers to the pair of drugs A and B
* Similarity Score: it denotes the similarity of fingerprints.
* MWrow/MWcol: . The Molecular weight of the molecule is indicated.
* LogP: he partition coefficients between octanol and water.
* NumHDonors: It indicates the quantity of hydrogen bond donors.
* NumHAcceptors: It indicates how many hydrogen bond acceptors there are

The target variable had to be defined after that. The target variable was chosen using similarity scores. Scores for similarity varied from 0 to 0.7. With "Synergistic" as the target term, we defined two classes. Our target variable has two classifications; thus, our issue is now a binary classification problem.

We classified synergy levels below 0.27 as non-synergistic or combative, and we assigned them a class of "0." Similarly, synergy values greater than 0.27 are regarded as synergistic, hence they belong to type class "1". We preprocessed the dataset after defining our target variable.

• The dataset's null values were first removed.

• Additionally, we removed the duplicated rows.

**B) Data Preprocessing**:

Data preprocessing is a crucial stage in machine learning since the quality of the data and the information that can be extracted from it directly influence how well our model can learn. As a result, it is important that we preprocess our data before feeding it to our model. We preprocessed the dataset after defining our target variable.

* The dataset's null values were first removed.
* Additionally, we removed the duplicated rows.
* We applied label encoding to drug row and drug col to integers because ml models cannot process the strings.
* Next, we created Synergistic scores using the similarity scores.

Chart, bar chart

Description automatically generated

Fig 3

**Results:**

Test/Train Split:

By giving the test size = 0.33 and random state = 10. For train and test, we divided the dataset.

Table

Description automatically generated

Fig 4

Test Results:

To find the best efficiency model, we used a variety of machine learning algorithms. We did the following

1. Random Forest
2. Decision Tree
3. XGBoost
4. Knn

We found that Random Forest and Decision Tree are our best performing model.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Accuracy | F1-Score | Cross Value Accuracy |
| Random Forest | 0.85 | 0.90 | 0.95 |
| Decision Tree | 0.85 | 0.88 | 0.94 |
| XGBoost | 0.57 | 0.7 | 0.36 |
| Knn | 0.57 | 0.4 | 0.56 |

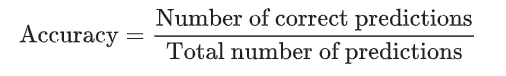
Table 3

For performance evaluation, we implemented evaluation metrics for the best model. Evaluating a model based on the output it produced by comparing it with original output is a core part of building an effective machine learning model.

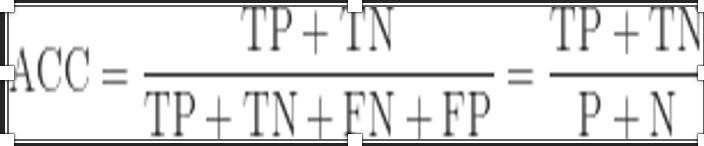
We checked for accuracy, f1\_macro, cross-validation score, and true positive rate.

For performance evaluation of the estimator, we computed cross-validation with scoring function as ‘f1-macro’ and “number of folds” = 5.

• Accuracy: One parameter for assessing classification models is accuracy. Accuracy is formally defined as the percentage of correct predictions generated by our model. The following is the official definition of accuracy:



Accuracy can also be determined in terms of positives and negatives for binary classification, as seen below:



• F1-score: To evaluate the efficacy of a test in a binary classification, we utilize the F1-score. The harmonic mean of the model's precision and recall is known as the F-score, which is a method of combining the model's precision and recall.

The F-score is frequently used to assess machine learning models of various kinds as well as information retrieval systems like search engines.



The best performing model was also plotted using a confusion matrix.

Chart

Description automatically generated

Fig 4

**Conclusion:**

One of the major issues facing by healthcare system is how to improve the effectiveness and personalization of cancer therapy. Combining different drugs may increase their ability to fight cancer drug resistance and, as a result, give patients more long-lasting therapy alternatives. Drug combination therapy is one potential method for treating diseases that are complex, such as cancer and other infectious disorders. Combining drugs is a promising strategy to increase treatment effectiveness, which is frequently limited by disease heterogeneity, toxicity, and drug resistance. However, it is impossible to compile a list of every combination that could have a synergistic effect. The project's goal is to build an efficient machine-learning model that finds effective drug combinations that are therapeutically applicable, to develop combinatorial drug treatments.

**References:**

[1] Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A., & Tang, J. (2019). DrugComb: an integrative cancer drug combination data portal. Nucleic Acids Research, 47(W1), W43–W51. https://doi.org/10.1093/nar/gkz337

[2] Lianlian Wu, Yuqi Wen, Dongjin Leng, Qinglong Zhang, Chong Dai, Zhongming Wang, Ziqi Liu, Bowei Yan, Yixin Zhang, Jing Wang, Song He, Xiaochen Bo (2021), Machine learning methods, databases and tools for drug combination prediction, *Briefings in Bioinformatics*, bbab355, <https://doi.org/10.1093/bib/bbab355>

[3] Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, Victor Robles (2006), Machine learning in bioinformatics, *Briefings in Bioinformatics*, 7(1), pp. 86–112, <https://doi.org/10.1093/bib/bbk007>

[4] Güvenç Paltun, B., Kaski, S., & Mamitsuka, H. (2021). Machine learning approaches for drug combination therapies. Briefings in Bioinformatics, 22(6). https://doi.org/10.1093/bib/bbab293

[5] Shipitsin, M., Polyak, K. The cancer stem cell hypothesis: in search of definitions, markers, and relevance. *Lab Invest* 88, 459–463 (2008). https://doi.org/10.1038/labinvest.2008.14

[6] Sung, H, Ferlay, J, Siegel, RL, Laversanne, M, Soerjomataram, I, Jemal, A, Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021: 71: 209- 249. <https://doi.org/10.3322/caac.21660>

[7] Funmilola A. Fisusi, Emmanuel O. Akala (2019), ‘Drug Combinations in Breast Cancer Therapy’, *Pharmaceutical Nanotechnology,* 7(1), pp. 3-23(21), 10.2174/2211738507666190122111224

[8] Li J, Tong X-Y, Zhu L-D and Zhang H-Y (2020) A Machine Learning Method for Drug Combination Prediction. Front. Genet. 11:1000. doi: 10.3389/fgene.2020.01000

[9] Ji X, Tong W, Liu Z and Shi T (2019) Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost. Front. Genet. 10:600. doi: 10.3389/fgene.2019.00600

[10] Lavecchia A. (2015), Machine-learning approaches in drug discovery: methods and applications, ‘*Drug Discovery Today*’, 20(3), pp. 318-331, <https://doi.org/10.1016/j.drudis.2014.10.012>

[11] Janizek, Joseph D., Celik, Safiye, Lee, Su-In (2018), Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine, ‘*bioRxiv*’, 331769, https://doi.org/10.1101/331769

[12] Benet LZ, Hosey CM, Ursu O, Oprea TI: BDDCS, the Rule of 5 and drugability. Adv Drug Deliv Rev. 2016 Jun 1;101:89-98. doi: 10.1016/j.addr.2016.05.007. Epub 2016 May 13.

[13] NEIDLE, and Stephen Neidle. (2012). Design Principles for Quadruplex-binding Small Molecules. In Therapeutic Applications of Quadruplex Nucleic Acids. Academic Press, <https://doi.org/10.1016/B978-0-12-375138-6.00009-1>

[14] Shuyu Zheng, Jehad Aldahdooh, Tolou Shadbahr, Yinyin Wang, Dalal Aldahdooh, Jie Bao, Wenyu Wang, Jing Tang, DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W174–W184, <https://doi.org/10.1093/nar/gkab438>

[15] Zhao, Wei, Kris Sachsenmeier, Lanju Zhang, Erin Sult, Robert E. Hollingsworth, and Harry Yang. “A New Bliss Independence Model to Analyze Drug Combination Data.” *Journal of Biomolecular Screening* 19, no. 5 (2014): 817–21. https://doi.org/10.1177/1087057114521867

11

[16] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001 Mar 1;46(1-3):3-26.

[17] Hirohara, M., Saito, Y., Koda, Y. *et al.* Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics* 19, 526 (2018). https://doi.org/10.1186/s12859-018-2523-5

[18] Liu, H., Zhang, W., Nie, L. *et al.* Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network. *BMC Bioinformatics* 20, 645 (2019). https://doi.org/10.1186/s12859-019-3288-1

[19] Lind AP, Anderson PC (2019) Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLoS ONE 14(7): e0219774. https://doi.org/10.1371/journal. pone.0219774

[20] Huang, H., Zhang, P., Qu, X. *et al.* Systematic prediction of drug combinations based on clinical side-effects. *Sci Rep* 4, 7160 (2014). https://doi.org/10.1038/srep07160

[21] SWINNEY, David C Swinney, and John E. E Macor. (2011). Molecular Mechanism of Action (MMoA) in Drug Discovery. In Annual Reports in Medicinal Chemistry. Academic Press, <https://doi.org/10.1016/B978-0-12-386009-5.00009-6>

[22] Yadav Bhagwan, Wennerberg Krister, Aittokallio Tero, Tang Jing, Searching for drug synergy in complex dose-response landscapes using an interaction potency model, Computational and Structural Biotechnology Journal (2015), doi: 10.1016/j.csbj.2015.09.00