Effective Targeted Drug Prediction for Cancer Based on Genetic Mutations

Darsh Mandera[1], Dr. Anna Ritz[2]

[1]Jesuit High School, [2]Reed College

Background

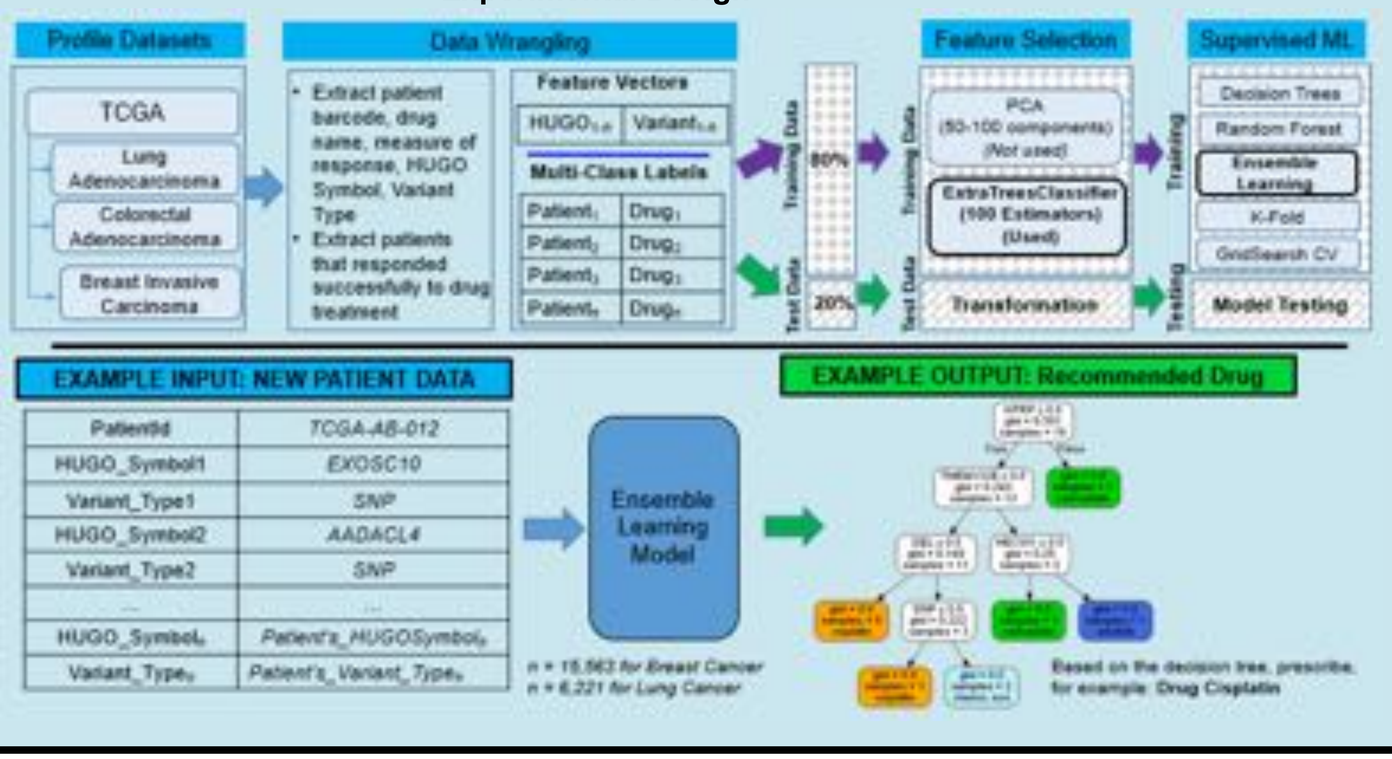
- Precision medicine aims to predict which treatment will be the most effective on a cancer patient based on their genetic mutations.
- Information about genetic changes in a tumor can help decide which treatment will work the best for a patient, resulting in a decrease in cancer morbidity.
- Today, prescribing a drug for cancer is a time-consuming process, as it is based on a doctor's exploratory research about a patient's genetic mutations and previous clinical trials.
- Machine learning has been used in various aspects of cancer prediction and prognosis. However, many of these studies use cancer cell lines, and not patient data.

Problem Statement: ***Create Machine Learning Classifier that Predicts which Drugs are Effective on a Cancer Patient with any Type of Cancer, based on their Genetic Mutations.***

Algorithms

Algorithm	Use of Algorithm
ExtraTrees Classifier	•Used for dimensionality reduction via feature selection. Selected the most important features (HUGO Symbols and Variant Types) •Iterative experimentation was done with ExtraTreesClassifier for reducing the dimensionality and overall accuracy was measured.
DecisionTree Classifier	The decision tree starts with a type of mutation and asks: *Does a patient have this mutation or not?* The tree then decides: *If not, then do not prescribe them any drug. If they do, prescribe Drug_X.*
AdaBoost Classifier	•RandomForestClassifier was the base estimator •Model training was performed on 100 estimators. •Works by combining multiple weak learning algorithms to create a strong learning algorithm.
OneVsRest Classifier	•Used to reduce the dimensionality of the dataset to ~3000 features to improve accuracy and accelerate training . •The algorithm selects one drug out of 20 drugs. Then combines the rest of the drugs into a single class. Next, it selects a certain combination of HUGO Symbols and Variant Types (out of 15,563) and decides whether it belongs in the one selected class or in the large combined class. The algorithm repeats this process until all possible permutations are executed
K-Nearest Neighbors Classifier	•Used to determine which patient should receive which drug. •Multiple values of k were tested, and ten returned the highest accuracy. •Uniform weightage was used in prediction •The output of this algorithm was which drug should be prescribed to which patient.
GridSearchCV	GridSearchCV is a hyperparameter optimization algorithm. It was used in this project to find the best training accuracy for a specific model.
K-Fold	Worked with GridSearchCV to cross-validate the training accuracy. *K* was selected to equal five (recommended best practice) because the number five is neither to small nor too big, resulting in efficiency and high accuracy.

Experimental Design



Results

Depth	OneVsRest Classifier (Ensemble Learning)	DecisionTree Classifier	AdaBoost Classifier (Ensemble Learning)
5	0.59195	0.75517	0.85356 (Best)
15	0.65242	0.67238	0.73679
25	0.78446	0.77346	0.63396
35	0.72397	0.70901	0.71837
45	0.6994	0.70275	0.77488
55	0.82821 (Best)	0.78792 (Best)	0.73919
65	0.74152	0.56026	0.81909
75	0.82698	0.73605	0.76644
85	0.70206	0.67905	0.72173
95	0.81606	0.65325	0.67563

Neighbors	*k*-NearestNeighbors Classifier
5	0.65759
6	0.62221
7	0.65633
8	0.63744
9	0.65606
10	0.68902 (Best)
11	0.63564
12	0.6415
13	0.65004
14	0.63938

Figure 1 (above): *Accuracy values for various machine learning algorithms on the BRCA dataset during the training phase.*

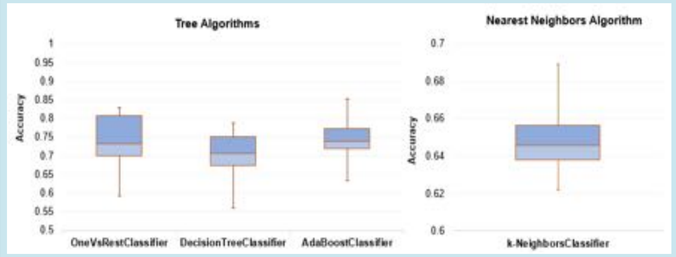


Figure 2 (left): *The variance of accuracy shows that a low sample size can result in a varying accuracy.*

	# Patients	# Drugs	Training Accuracy	Testing Accuracy
Breast Invasive Carcinoma (BRCA)	982	104	83%	66%
Lung Adenocarcinoma (LUAD)	178	33	55%	50%
Colorectal Adenocarcinoma (COAD)	223	23	66%	66%

Figure 3 (above): *The final approximate training and testing accuracy values for all three cancer types that were tested.*

Conclusions and Future Directions

•Conclusion: The program can be used by doctors to identify which drugs to prescribe to a cancer patient for targeted therapy, and it works on all cancer types.

•Future Direction: Test the program against more cancer types; expand the program to use more dimensions, like patients' lifestyles, to predict with a higher accuracy, and/or create a new program that predicts partial or complete efficacy of a more than one drug; collaborate with a pharmaceutical company to pilot the program in cancer drug trials.