An Ensemble Learning Approach for Cancer Drug Prediction

Darsh Mandera[1] and Anna Ritz[2]

[1] Jesuit High School, Portland, Oregon, USA; dsmandera@gmail.com
[2] Reed College, Portland, Oregon, USA; aritz@reed.edu

Abstract

Predicting the response to a particular drug for specific cancer, despite known genetic mutations, still remains a huge challenge in modern oncology and precision medicine. Today, prescribing a drug for a cancer patient is based on a doctor's analysis of various articles and previous clinical trials; it is an extremely time-consuming process. We developed a machine learning classifier to automatically predict a drug given a carcinogenic gene mutation profile. Using the Breast Invasive Carcinoma Dataset from The Cancer Genome Atlas (TCGA), the method first selects features from mutated genes and then applies K-Fold, Decision Tree, Random Forest and Ensemble Learning classifiers to predict best drugs. Ensemble Learning yielded prediction accuracy of 66% on the test set in predicting the correct drug. To validate that the model is general-purpose, Lung Adenocarcinoma (LUAD) data and Colorectal Adenocarcinoma (COADREAD) data from TCGA was trained and tested, yielding prediction accuracies 50% and 66% respectively. The resulting accuracy indicates a direct correlation between prediction accuracy and cancer data size. More importantly, the results of LUAD and COADREAD show that the implemented model is general purpose as it is able to achieve similar results across multiple cancer types. We further verified the validity of the model by implementing it on patients with unclear recovery status from the COADREAD dataset. In every case, the model predicted a drug that was administered to each patient. This method will offer oncologists significant time-saving compared to their current approach of extensive background research, and offers personalized patient care for cancer patients.

Keywords: Machine learning, Cancer drug prediction, Breast Invasive Carcinoma (BRCA), Lung Adenocarcinoma (LUAD), Colorectal Adenocarcinoma (COADREAD), The Cancer Genome Atlas (TCGA), gene mutation, ensemble learning

Introduction

Almost one in six deaths occur due to cancer, and the number of new cancer cases is expected to rise by 70% in the next twenty years ("WHO | Cancer", 2017). In today's world, the process of prescribing a drug for a cancer patient is based on a doctor's exploratory research through various scientific journals, such as the Cochrane Review, PubMed, et cetera, as well as previous clinical trials. This can become an extremely time-consuming and iterative process.

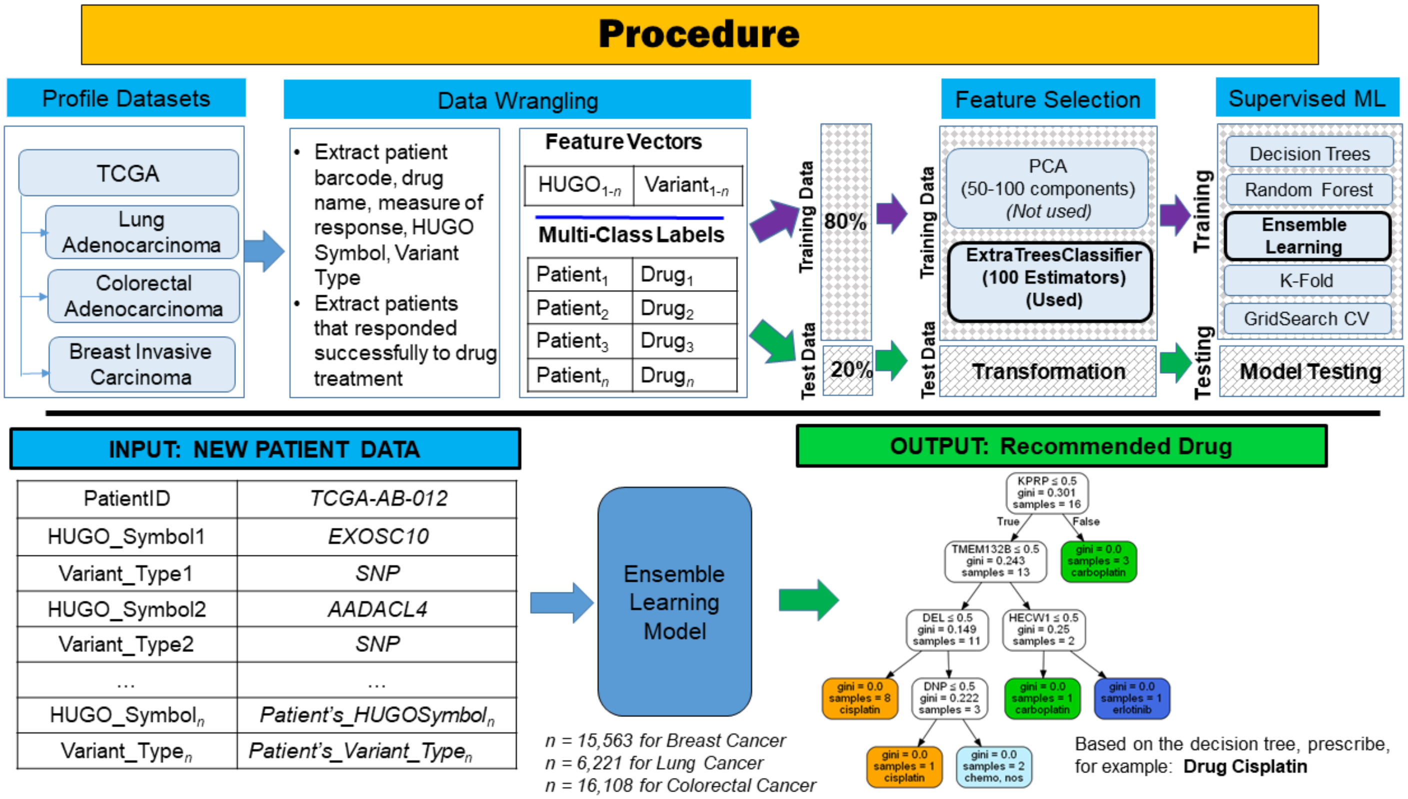
The principal goal of precision medicine is to treat patients with the treatments that are most likely to help the patients, based on their genome ("Precision Medicine in Cancer Treatment – National Cancer Institute", 2017). In cancer, this entails predicting which treatment will be the most effective on a patient based on their genetic mutations. Machine learning, which is a branch of artificial intelligence that allows computers to "learn" from experience, has been used in various aspects of cancer prediction and prognosis (Simes, 1985; Maclin et al., 1991; Ciccheti, 1992; Petricoin and Liotta, 2004; Bocchi et al., 2004; Zhou et al., 2004; Dettling, 2004; McCarthy et al., 2004; Wang et al., 2005; Cruz and Wishart, 2007, Yasrebi et al., 2009, Yasrebi, 2016). Combined with a high volume and wide range of patients' records consisting of gene data along with effectiveness of medications, machine learning will help predict the best way to administer the correct targeted medication for patients with cancer (Kourou et al., 2014; Bashiri et al., 2017). In today's world, the majority of such established machine learning programs use data from cancer cell lines – sets of cancer cells that grow and divide in a laboratory. Still others use humanized mice models – human xenografts and immune cells grafted into lab mice (Kalamara, et el., 2018). Although these may be effective models, it is impossible for any model to fully capture the complexity of real patient data.

This project utilizes supervised machine learning and aims at creating a general-purpose machine learning classifier that can predict which drug will be effective for a cancer patient based on their genetic mutations, for any cancer type. Overall, this program uses patterns and correlations between a patient's response to a drug and their genetic mutations in order to predict which drug will be effective on a given cancer patient, based on their genetic mutations.

Methods

This project involved multiple steps including data wrangling, training the machine learning model, and testing the machine learning model (Figure 1).

The first stage of this project was data acquisition and data wrangling. The data was acquired from The Cancer Genome Atlas (TCGA), which is a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), and has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer ("The Cancer Genome Atlas Home Page," n.d.). First, a Breast Invasive Carcinoma (BRCA) dataset with 982 patients was downloaded. Patient barcode (patient's anonymous TCGA ID), drug names, measure of response (a patient's response to a specific drug), HUGO Symbol (name of the gene where a genetic mutation occurs), and variant type (the type of genetic mutation that a patient has) were extracted from the BRCA dataset consisting of over 15,000 data types (features).


Procedure
Profile Datasets
TCGA
Lung Adenocarcinoma
Colorectal Adenocarcinoma
Breast Invasive Carcinoma
Data Wrangling
Extract patient barcode, drug name, measure of response, HUGO Symbol, Variant Type
Extract patients that responded successfully to drug treatment
Feature Vectors
$HUGO_{1-n}$
$Variant_{1-n}$
Multi-Class Labels
$Patient_1$ $Drug_1$
$Patient_2$ $Drug_2$
$Patient_3$ $Drug_3$
$Patient_n$ $Drug_n$
Training Data
80%
Test Data
20%
Feature Selection
Training Data
PCA (50-100 components) (Not used)
ExtraTreesClassifier (100 Estimators) (Used)
Test Data
Transformation
Supervised ML
Training
Decision Trees
Random Forest
Ensemble Learning
K-Fold
GridSearch CV
Testing
Model Testing
INPUT: NEW PATIENT DATA
Ensemble Learning Model
OUTPUT: Recommended Drug
KPRP ≤ 0.5 gini = 0.301 samples = 16
True
False
TMEM132B ≤ 0.5 gini = 0.243 samples = 13
gini = 0.0 samples = 3 carboplatin
DEL ≤ 0.5 gini = 0.149 samples = 11
HECW1 ≤ 0.5 gini = 0.25 samples = 2
gini = 0.0 samples = 8 cisplatin
DNP ≤ 0.5 gini = 0.222 samples = 3
gini = 0.0 samples = 1 carboplatin
gini = 0.0 samples = 1 erlotinib
gini = 0.0 samples = 1 cisplatin
gini = 0.0 samples = 2 chemo, nos

PatientID	*TCGA-AB-012*
HUGO_Symbol1	*EXOSC10*
Variant_Type1	*SNP*
HUGO_Symbol2	*AADACL4*
Variant_Type2	*SNP*
…	…
$HUGO\_Symbol_n$	*$Patient's\_HUGOSymbol_n$*
$Variant\_Type_n$	*$Patient's\_Variant\_Type_n$*

n = 15,563 for Breast Cancer
n = 6,221 for Lung Cancer
n = 16,108 for Colorectal Cancer

Based on the decision tree, prescribe, for example: **Drug Cisplatin**

Figure 1: Various stages of the machine learning model training and prediction on various cancer datasets. Data from The Cancer Genome Atlas were formatted to extract features (mutations and variants) and multi-class drug labels for each patient. Then, we performed feature selection and supervised machine learning for four different classifiers. Once the model is trained, new patient data can be classified in the Ensemble Learning Model to recommend a drug for that patient.

In machine learning, feature selection can be undertaken in order to remove features that do not contribute to the accuracy of the model. In this research, we implemented feature selection using the ExtraTreesClassifier on the BRCA dataset. The ExtraTreesClassifier implements an estimator that fits randomized decision trees on samples of the dataset in order to select features that have the highest impact on the final accuracy, allowing for reduction of overfitting. ("3.2.4.3.3. sklearn.ensemble.ExtraTreesClassifier", 2019) Principal Component Analysis (PCA) was also tested for feature selection, but ExtraTreesClassifier resulted in a higher accuracy of 52%, while PCA returned an accuracy of 30%. The use of the ExtraTreesClassifier resulted in the identification of the top 20 features (various HUGO Symbols and Variant Types), and their impact on the accuracy of the program (Figure 2).

The resulting BRCA dataset had 695 patients that had successfully responded to drugs (i.e. had a complete response) for their particular genetic mutation profile. For patients who were administered more than one drug, we chose the first drug that resulted in a complete response as the target. This dataset was then converted into two vectors: one containing all HUGO Symbols and Variant Types, and the other containing patient barcodes and drug names.

Once the data had been curated, the extracted dataset of 695 patients was split into a training section and a testing section using the train_test_split function in scikit-learn. 80% of the data was used for training, and the remaining 20% was used for testing, as an 80-20 split is a recommended best practice.

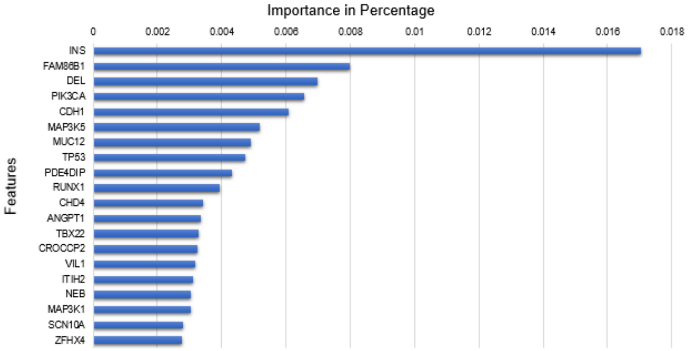


Figure 2: Top 20 features that were selected from the BRCA dataset.

The training dataset was used to fit the model to learn which drug is the most effective for a set of mutated genes. On the training data with features, K-Fold and GridSearchCV were implemented for cross-validation. K-Fold, with a K-value selected to equal 5, split the training dataset into five equal folds. GridSearchCV was then implemented on each of these five folds as a hyperparameter optimization algorithm and selected the best parameters to use for prediction. The processes of train_test_split and cross-validation were repeated using four different classifiers, in order to identify which machine learning algorithms would result in the highest prediction accuracy. The following is a brief description of the four algorithms that we tested:

DecisionTree Classifier: A decision tree outputs a tree-like structure that decides which drug a patient should receive based on their genetic mutations. 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.* ("1.10 Decision Trees", 2020)

AdaBoost Classifier: The AdaBoostClassifier combines weak algorithms to form a strong prediction algorithm. AdaBoostClassifier works by training multiple base estimators on the training data, and then using a weighted voter among those estimators in order to select a final classification. In this study, we chose our base estimator to be RandomForest. ("sklearn.ensemble.AdaBoostClassifier", 2019)

One vs. Rest Classifier: This classifier takes a unique approach on classifying an item. The program selects one drug, then combines the rest of the drugs into a single class. Then, it selects a certain combination of HUGO Symbols and Variant Types and decides whether the combination belongs in the one selected class or in the large combined class. Then, the algorithm selects a new drug to isolate. The algorithm repeats this process until all possible permutations are executed. ("sklearn.multiclass.OneVsRestClassifier", 2020)

K-Nearest-Neighbor Classifier: This last algorithm works by learning from *k* closest neighbors to an unclassified patient. It then chooses a classification for the given patient based on the class of its *k* closest neighbors. ("1.6. Nearest Neighbors", 2019).

In order to validate that the final machine learning model was general-purpose, a Lung Adenocarcinoma (LUAD) dataset with 178 patients and 33 drugs and a Colorectal Adenocarcinoma (COADREAD) dataset with 223 patients and 23 drugs were downloaded from TCGA, and these datasets underwent a similar data wrangling process as the BRCA dataset. LUAD and COADREAD data was then inputted into the machine learning model. The model was designed and implemented in Python 3.6 using scikit-learn 0.19.1, pandas 0.20.3, and pydot 1.2.3 in order to predict the best drug for all cancer types. All code and data are available at https://github.com/annaritz/cancer-drug-pred.

Results

Overall, AdaBoostClassifier and OneVsRestClassifier yielded two of the highest training prediction accuracies of 85% and 83% respectively, and they were implemented in the testing phase. The BRCA test data of 139 patients was scored on the trained model based on AdaBoostClassifier and OneVsRestClassifier. AdaBoostClassifier yielded an accuracy of 63%, while OneVsRestClassifier yielded the highest testing accuracy of 66%. Therefore, OneVsRestClassifier was used in the final model.

The final machine learning model was tested on BRCA data as well as LUAD and COADREAD datasets to evaluate it as a general-purpose program. On the BRCA data, the program was able to predict effective drugs on a patient with a training accuracy of 83% and a testing accuracy of 66%. On the LUAD dataset, the program achieved a training accuracy of 55% and a testing accuracy of 50%. On the COADREAD dataset, the training accuracy and testing accuracy were both 66%. The reason that the accuracy values vary highly is due to the fact that some datasets (e.g. LUAD) may be sparse or have a smaller sample size compared to other datasets. (Figure 3).

In this study, the most commonly selected drugs for Breast Invasive Carcinoma were cyclophosphamide (sometimes labeled as Cytoxan), tamoxifen, doxorubicin (labeled as Adriamycin), and anastrozole (labeled as Arimidex). The DecisionTree Classifier for BRCA is shown in Figure 4. Cyclophosphamide is often used to treat all types of breast cancer, in addition to other cancers like ovarian cancer and leukemia. As an alkylating agent, cyclophosphamide slows the growth of breast cancer cells or, in some cases, completely halts the growth. ("Cyclophosphamide: MedicinePlus Drug Information", 2018; Cocquyt et al., 2003). Tamoxifen, another commonly selected drug for BRCA, works by binding to estrogen receptors in order to prevent estrogen binding ("Hormone Therapy for Breast Cancer Fact Sheet - National Cancer Institute", 2017). It too has been used to treat breast cancer (Baum et al., 2002; Houghton et al., 2003; Silva et al., 2014). Doxorubicin and anastrozole both work by slowing down cancer cell growth; the former by blocking topo isomorase 2 and the latter by blocking estrogen receptors ("Doxorubicin (Adryiamycin) | Cancer Drugs | Cancer Research UK", 2017; "Anastrozole", 2018). Previous research has demonstrated these drugs' efficacy in BRCA patients (Baselga et al., 1998; Baum et al., 2002; Chae et al, 2016).

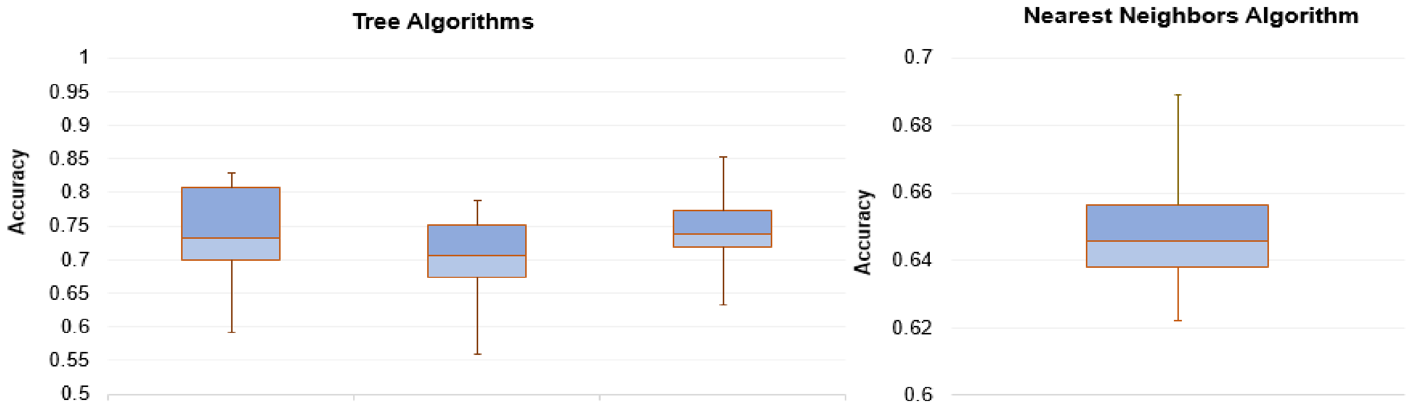


Figure 3: The variance of accuracy of the program tested on the BRCA dataset.

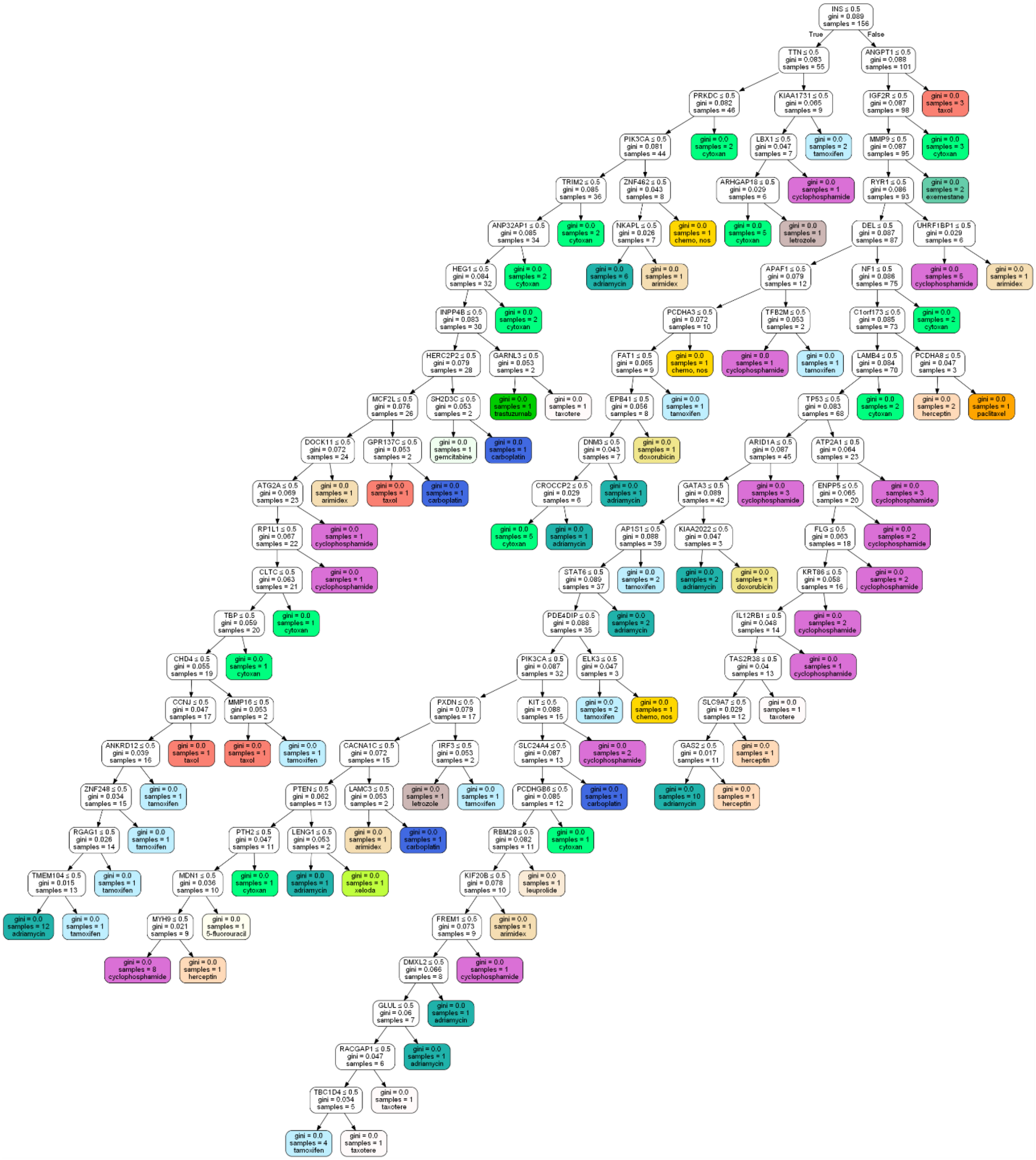

INS ≤ 0.5 gini = 0.089 samples = 156
True
False
TTN ≤ 0.5 gini = 0.083 samples = 55
ANGPT1 ≤ 0.5 gini = 0.088 samples = 101
PRKDC ≤ 0.5 gini = 0.082 samples = 46
KIAA1731 ≤ 0.5 gini = 0.065 samples = 9
IGF2R ≤ 0.5 gini = 0.087 samples = 98
gini = 0.0 samples = 3 taxol
PIK3CA ≤ 0.5 gini = 0.081 samples = 44
gini = 0.0 samples = 2 cytoxan
LBX1 ≤ 0.5 gini = 0.047 samples = 7
gini = 0.0 samples = 2 tamoxifen
MMP9 ≤ 0.5 gini = 0.087 samples = 95
gini = 0.0 samples = 3 cytoxan
TRIM2 ≤ 0.5 gini = 0.085 samples = 36
ZNF462 ≤ 0.5 gini = 0.043 samples = 8
ARHGAP18 ≤ 0.5 gini = 0.029 samples = 6
gini = 0.0 samples = 1 cyclophosphamide
RYR1 ≤ 0.5 gini = 0.086 samples = 93
gini = 0.0 samples = 2 exemestane
ANP32AP1 ≤ 0.5 gini = 0.085 samples = 34
gini = 0.0 samples = 2 cytoxan
NKAPL ≤ 0.5 gini = 0.026 samples = 7
gini = 0.0 samples = 1 chemo, nos
gini = 0.0 samples = 5 cytoxan
gini = 0.0 samples = 1 letrozole
DEL ≤ 0.5 gini = 0.087 samples = 87
UHRF1BP1 ≤ 0.5 gini = 0.029 samples = 6
HEG1 ≤ 0.5 gini = 0.084 samples = 32
gini = 0.0 samples = 2 cytoxan
gini = 0.0 samples = 6 adriamycin
gini = 0.0 samples = 1 arimidex
APAF1 ≤ 0.5 gini = 0.079 samples = 12
NF1 ≤ 0.5 gini = 0.086 samples = 75
gini = 0.0 samples = 5 cyclophosphamide
gini = 0.0 samples = 1 arimidex
INPP4B ≤ 0.5 gini = 0.083 samples = 30
gini = 0.0 samples = 2 cytoxan
PCDHA3 ≤ 0.5 gini = 0.072 samples = 10
TFB2M ≤ 0.5 gini = 0.053 samples = 2
C1orf173 ≤ 0.5 gini = 0.085 samples = 73
gini = 0.0 samples = 2 cytoxan
HERC2P2 ≤ 0.5 gini = 0.079 samples = 28
GARNL3 ≤ 0.5 gini = 0.053 samples = 2
FAT1 ≤ 0.5 gini = 0.065 samples = 9
gini = 0.0 samples = 1 chemo, nos
gini = 0.0 samples = 1 cyclophosphamide
gini = 0.0 samples = 1 tamoxifen
LAMB4 ≤ 0.5 gini = 0.084 samples = 70
PCDHA8 ≤ 0.5 gini = 0.047 samples = 3
MCF2L ≤ 0.5 gini = 0.076 samples = 26
SH2D3C ≤ 0.5 gini = 0.053 samples = 2
gini = 0.0 samples = 1 trastuzumab
gini = 0.0 samples = 1 taxotere
EPB41 ≤ 0.5 gini = 0.056 samples = 8
gini = 0.0 samples = 1 tamoxifen
TP53 ≤ 0.5 gini = 0.083 samples = 68
gini = 0.0 samples = 2 cytoxan
gini = 0.0 samples = 2 herceptin
gini = 0.0 samples = 1 paclitaxel
DOCK11 ≤ 0.5 gini = 0.072 samples = 24
GPR137C ≤ 0.5 gini = 0.053 samples = 2
gini = 0.0 samples = 1 gemcitabine
gini = 0.0 samples = 1 carboplatin
DNM3 ≤ 0.5 gini = 0.043 samples = 7
gini = 0.0 samples = 1 doxorubicin
ARID1A ≤ 0.5 gini = 0.087 samples = 45
ATP2A1 ≤ 0.5 gini = 0.064 samples = 23
ATG2A ≤ 0.5 gini = 0.069 samples = 23
gini = 0.0 samples = 1 arimidex
gini = 0.0 samples = 1 taxol
gini = 0.0 samples = 1 carboplatin
CROCCP2 ≤ 0.5 gini = 0.029 samples = 6
gini = 0.0 samples = 1 adriamycin
GATA3 ≤ 0.5 gini = 0.089 samples = 42
gini = 0.0 samples = 3 cyclophosphamide
ENPP5 ≤ 0.5 gini = 0.065 samples = 20
gini = 0.0 samples = 3 cyclophosphamide
RP1L1 ≤ 0.5 gini = 0.067 samples = 22
gini = 0.0 samples = 1 cyclophosphamide
gini = 0.0 samples = 5 cytoxan
gini = 0.0 samples = 1 adriamycin
AP1S1 ≤ 0.5 gini = 0.088 samples = 39
KIAA2022 ≤ 0.5 gini = 0.047 samples = 3
FLG ≤ 0.5 gini = 0.063 samples = 18
gini = 0.0 samples = 2 cyclophosphamide
CLTC ≤ 0.5 gini = 0.063 samples = 21
gini = 0.0 samples = 1 cyclophosphamide
STAT6 ≤ 0.5 gini = 0.089 samples = 37
gini = 0.0 samples = 2 tamoxifen
gini = 0.0 samples = 2 adriamycin
gini = 0.0 samples = 1 doxorubicin
KRT86 ≤ 0.5 gini = 0.058 samples = 16
gini = 0.0 samples = 2 cyclophosphamide
TBP ≤ 0.5 gini = 0.059 samples = 20
gini = 0.0 samples = 1 cytoxan
PDE4DIP ≤ 0.5 gini = 0.088 samples = 35
gini = 0.0 samples = 2 adriamycin
IL12RB1 ≤ 0.5 gini = 0.048 samples = 14
gini = 0.0 samples = 2 cyclophosphamide
CHD4 ≤ 0.5 gini = 0.055 samples = 19
gini = 0.0 samples = 1 cytoxan
PIK3CA ≤ 0.5 gini = 0.087 samples = 32
ELK3 ≤ 0.5 gini = 0.047 samples = 3
TAS2R38 ≤ 0.5 gini = 0.04 samples = 13
gini = 0.0 samples = 1 cyclophosphamide
CCNJ ≤ 0.5 gini = 0.047 samples = 17
MMP16 ≤ 0.5 gini = 0.053 samples = 2
PXDN ≤ 0.5 gini = 0.079 samples = 17
KIT ≤ 0.5 gini = 0.088 samples = 15
gini = 0.0 samples = 2 tamoxifen
gini = 0.0 samples = 1 chemo, nos
SLC9A7 ≤ 0.5 gini = 0.029 samples = 12
gini = 0.0 samples = 1 taxotere
ANKRD12 ≤ 0.5 gini = 0.039 samples = 16
gini = 0.0 samples = 1 taxol
gini = 0.0 samples = 1 taxol
gini = 0.0 samples = 1 tamoxifen
CACNA1C ≤ 0.5 gini = 0.072 samples = 15
IRF3 ≤ 0.5 gini = 0.053 samples = 2
SLC24A4 ≤ 0.5 gini = 0.087 samples = 13
gini = 0.0 samples = 2 cyclophosphamide
GAS2 ≤ 0.5 gini = 0.017 samples = 11
gini = 0.0 samples = 1 herceptin
ZNF248 ≤ 0.5 gini = 0.034 samples = 15
gini = 0.0 samples = 1 tamoxifen
PTEN ≤ 0.5 gini = 0.062 samples = 13
LAMC3 ≤ 0.5 gini = 0.053 samples = 2
gini = 0.0 samples = 1 letrozole
gini = 0.0 samples = 1 tamoxifen
PCDHGB6 ≤ 0.5 gini = 0.085 samples = 12
gini = 0.0 samples = 1 carboplatin
gini = 0.0 samples = 10 adriamycin
gini = 0.0 samples = 1 herceptin
RGAG1 ≤ 0.5 gini = 0.026 samples = 14
gini = 0.0 samples = 1 tamoxifen
PTH2 ≤ 0.5 gini = 0.047 samples = 11
LENG1 ≤ 0.5 gini = 0.053 samples = 2
gini = 0.0 samples = 1 arimidex
gini = 0.0 samples = 1 carboplatin
RBM28 ≤ 0.5 gini = 0.082 samples = 11
gini = 0.0 samples = 1 cytoxan
TMEM104 ≤ 0.5 gini = 0.015 samples = 13
gini = 0.0 samples = 1 tamoxifen
MDN1 ≤ 0.5 gini = 0.036 samples = 10
gini = 0.0 samples = 1 cytoxan
gini = 0.0 samples = 1 adriamycin
gini = 0.0 samples = 1 xeloda
KIF20B ≤ 0.5 gini = 0.078 samples = 10
gini = 0.0 samples = 1 leuprolide
gini = 0.0 samples = 12 adriamycin
gini = 0.0 samples = 1 tamoxifen
MYH9 ≤ 0.5 gini = 0.021 samples = 9
gini = 0.0 samples = 1 5-fluorouracil
FREM1 ≤ 0.5 gini = 0.073 samples = 9
gini = 0.0 samples = 1 arimidex
gini = 0.0 samples = 8 cyclophosphamide
gini = 0.0 samples = 1 herceptin
DMXL2 ≤ 0.5 gini = 0.066 samples = 8
gini = 0.0 samples = 1 cyclophosphamide
GLUL ≤ 0.5 gini = 0.06 samples = 7
gini = 0.0 samples = 1 adriamycin
RACGAP1 ≤ 0.5 gini = 0.047 samples = 6
gini = 0.0 samples = 1 adriamycin
TBC1D4 ≤ 0.5 gini = 0.034 samples = 5
gini = 0.0 samples = 1 taxotere
gini = 0.0 samples = 4 tamoxifen
gini = 0.0 samples = 1 taxotere

Figure 4: A decision tree generated based on the BRCA dataset that determines which drug a patient should receive based on their genetic mutations.

In order to further test the validity of the classifier, 200 patients from the COADREAD dataset who did not recover from their disease were selected. These patients varied in age and sex. Fifty-two percent of these unclassified patients were male, while the remaining 48% were female. Two percent were under 18, 2% were between the ages of 19 and 40, 15% were between the ages of 41 and 64, and 81% were over 65. For each of these patients, the classifier predicted a drug that had been administered by the clinic but did not help the patient recover. For example, for the patient TCGA-A6-2677, the classifier predicted to prescribe the drug oxaliplatin, which had been administered by the clinic. The most commonly selected drugs for unclassified patients were oxaliplatin, 5-fluorouracil, and bevacizumab, which are among the most commonly prescribed drugs for colorectal adenocarcinoma (“Targeted Therapy Drugs for Colorectal Cancer”, 2018; “Chemotherapy for Colorectal Cancer”, 2018). Overall, these results show that the classifier is able to predict drugs that are commonly prescribed by oncologists in clinical settings.

Discussion

Predicting which drugs will be most effective on a cancer patient based on their genetic mutations is an important goal of precision medicine. A general-purpose machine learning classifier was designed and implemented to predict targeted drug for cancer based on patients’ genetic mutations. Overall, the program that was created can be used by doctors to identify which drugs to prescribe to a cancer patient for targeted therapy. While the model is able to predict drugs similarly to a doctor’s judgement, it is important to note that the current tool still cannot distinguish between when drugs will be effective and when patients will not respond.

The next steps to this project will be to further validate the general purpose of the machine learning model by testing it against other cancer types (e.g. Head and Neck, Prostrate, et cetera), ideally with larger datasets. Also, expanding the machine learning model to use more dimensions such as patient lifestyle and location of the mutated gene to narrow the scope and predict with higher accuracy (90% or more) would be ideal. Another plan could be to create a new machine learning model that predicts partial or complete efficacy of more than one drug, that is, model the program as a multi-class problem.

Author Contributions

D.M. began this project in 2017, and sought the mentorship of A.R. D.M. and A.R. formulated the prediction problem, and D.M. carried out all experiments. D.M. wrote the initial drafts of the manuscript, and D.M. and A.R. worked on the manuscript together. D.M. also presented this research at the Reed College Poster Session as well as the Oregon Health and Science University Research Week in 2018, and the Rocky Mountains Bioinformatics Conference in 2019.

References

1. 1.10. Decision Trees. (2020). Retrieved from https://scikit-learn.org/stable/modules/tree.html
2. 1.6. Nearest Neighbors. (2019). Retrieved from https://scikit-learn.org/stable/modules/neighbors.html
3. 3.2.4.3.3. sklearn.ensemble.ExtraTreesClassifier. (2019). Retrieved August 03, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
4. Baselga, J., Norton, L., Albanell, J., Kim, Y., & Mendelsohn, J. (1998, July 01). Recombinant Humanized Anti-HER2 Antibody (Herceptin™) Enhances the Antitumor Activity of Paclitaxel and Doxorubicin against HER2/neu Overexpressing Human Breast Cancer Xenografts. Retrieved from https://cancerres.aacrjournals.org/content/58/13/2825.short
5. Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017, February). Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. Retrieved from https://ncbi.nlm.nih.gov/pmc/articles/PMC5402773/
6. Baum, M., Budzar, A. U., Cuzick, J., Forbes, J., Houghton, J. H., Klijn, J. G., & Sahmoud, T. (2002, November 9). Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: First results of the ATAC randomised trial. Retrieved from https://pubmed.ncbi.nlm.nih.gov/12090977/
7. Bocci, L., Coppini, G., Nori, J., & Valli, G. (2004, May). Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. Retrieved from https://medengphys.com/article/S1350-4533(03)00208-X/fulltext
8. Chae, J., Ng, T., Yeo, H. L., Shwe, M., Gan, Y. X., & Chan, A. (2016). Impact of TNF-α (rs1800629) and IL-6 (rs1800795) Polymorphisms on Cognitive Impairment in Asian Breast Cancer Patients. *Plos One, 11*(10). doi:10.1371/journal.pone.0164204
9. Cicchetti, D. V. (1992, January 01). Neural networks and diagnosis in the clinical laboratory: State of the art. Retrieved from http://clinchem.aaccjnls.org/content/38/1/9.long
10. Cocquyt, V. F., Blondeel, P. N., Depypere, H. T., Praet, M. M., Schelfhout, V. R., Silva, O. E., . . . Van Belle, S. J. (2003). Different responses to preoperative chemotherapy for invasive lobular and invasive ductal breast carcinoma. Retrieved from https://pubmed.ncbi.nlm.nih.gov/12711290/
11. Cruz, J. A., & Wishart, D. S. (2007, February 11). Applications of Machine Learning in Cancer Prediction and Prognosis. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/
12. Cyclophosphamide: MedlinePlus Drug Information. (2018). Retrieved from https://medlineplus.gov/druginfo/meds/a682080.html
13. Dettling, M. (2004, October 05). BagBoosting for tumor classification with gene expression data | Bioinformatics | Oxford Academic. Retrieved from https://academic.oup.com/bioinformatics/article/20/18/3583/20253
14. Hormone Therapy for Breast Cancer Fact Sheet. (2017). Retrieved from https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet
15. Houghton, J., Ioffe, O., Silverberg, S., McGrady, B., & McCluggage, W. (2003, April 01). Metastatic Breast Lobular Carcinoma Involving Tamoxifen-Associated Endometrial Polyps: Report of Two Cases and Review of Tamoxifen-Associated Polypoid Uterine Lesions. Retrieved from https://www.nature.com/articles/3880777
16. Kalamara, A., Tobalina, L., & Saez-Rodriguez, J. (2018, July 12). How to find the right drug for each patient? Advances and challenges in pharmacogenomics. Retrieved from https://www.sciencedirect.com/science/article/pii/S2452310018300076.
17. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014, November 15). Machine learning applications in cancer prognosis and prediction. Retrieved from https://www.sciencedirect.com/science/article/pii/S2001037014000464
18. Maclin, P. S., Dempsey, J., Brooks, J., & Rand, J. (1991, February). Using neural networks to diagnose cancer. Retrieved from https://ncbi.nlm.nih.gov/pubmed/1748845
19. McCarthy, J. F., Marx, K. A., Hoffman, P. E., Gee, A. G., O'Neil, P., Ujwal, M. L., & Hotchkiss, J. (2006, January 12). Applications of

Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis, and Management. Retrieved from https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1196/annals.1310.020

20. Petricoin, E. F., & Liotta, L. A. (2004, January 23). SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. Retrieved from https://sciencedirect.com/science/article/pii/S0958166904000072?via=ihub

21. Precision Medicine in Cancer Treatment. (2017, October 3). Retrieved from https://www.cancer.gov/about-cancer/treatment/types/precision-medicine

22. Simes, R. J. (1985). Treatment selection for cancer patients: Application of statistical decision theory to the treatment of advanced ovarian cancer. Retrieved from https://ncbi.nlm.nih.gov/pubmed/3882734

23. sklearn.ensemble.AdaBoostClassifier. (2019). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

24. sklearn.multiclass.OneVsRestClassifier. (2020). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

25. The Cancer Genome Atlas Home Page. (n.d.). Retrieved from https://cancergenome.nih.gov/

26. Wang, J. X., Zhang, B., Yu, J. K., Liu, J., Yang, M. Q., & Zheng, S. (2005, August 05). Application of serum protein fingerprinting coupled with artificial neural network model in diagnosis of hepatocellular carcinoma. Retrieved from https://ncbi.nlm.nih.gov/pubmed/16117882

27. WHO | Cancer. (2017). Retrieved November 21, 2017, from http://www.who.int/mediacentre/factsheets/fs297/en/

28. Yasrebi, H., Sperisen, P., Praz, V., & Bucher, P. (2009). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2761544/

29. Yasrebi, H. (2016, September). Comparative study of joint analysis of microarray gene expression data in survival prediction and risk assessment of breast cancer patients. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863785/

30. Zhou, X., Liu, K., & Wong, S. T. C. (2004, September 11). Cancer classification and prediction using logistic regression with Bayesian gene selection. Retrieved from https://sciencedirect.com/science/article/pii/S1532046404000772?via=ihub