Target Gene Identification Through CRISPR KO Labelling

Tanishq Bhatia
Northeastern University

Yuyang Cao
Northeastern University

Jieying Jin
Northeastern University

Jundong Wang
Northeastern University

Jiaming Xu
Northeastern University

Xueqi Xi
Northeastern University

Abstract

A. Bazaga's paper provides a machine learning based methodology for identifying potential target genes for cancer drugs. We wish to replicate their methodology in this paper and to extend the methodlogy to identify potential target genes for CRISPR KO. In this paper, we apply six different machine learning models to the DepMap dataset towards this end and interpret our results from the perspective of the original paper.

1 Introduction

Suboptimal target selection will cause costly development of a modulator that is ineffective at treating the pathology of interest. Under this motivation, people in drug development area participated themselves into investigation of target selection and validation. Nowadays Computational intelligence methods are introduced and involved into the investigation. In this paper authors tried applying 6 different machine learning classifiers on data set obtain from public resource to work out drug target classification for nine different human cancer types which includes bladder, breast, kidney, colon, leukemia, liver, lung, ovarian and pancreatic.

For each cancer type and each method the model is trained by data set contains a set of "know" target genes extended with equally-sized sets of "non-targets" on three primary features, mutation, gene expression and gene essentiality (DepMap) data extended with a set of features selected based on features importance. And as a result, they ran prediction on more than 15,000 protein-coding genes to identify potential novel targets.

We tried KNN method, Artificial Neural Network, Support Vector Machine, Linear Discriminant Analysis, Random Forest, Logistic Regression and represent the (Receiver operating characteristic (ROC) graph for each one and used ROC and AUC to studying the performance of each method on each cancer type.

1.1 Concept of CRISPR KO

In the CRISPR KO strategy, a CAS 9 Nuclease generates a double-stranded break on the targeted genomic site guided by a single RNA identified as an on-target site. The breaks on the DNA are later repaired by the cell's NHEJ (non-homologous end joining) repair mechanism that will trigger a site-specific InDels (Insertion/deletions), which will affect the gene's activity. The site where the break should be generated is typically decided by specific software. In case of the site is located

Preprint. Under review.

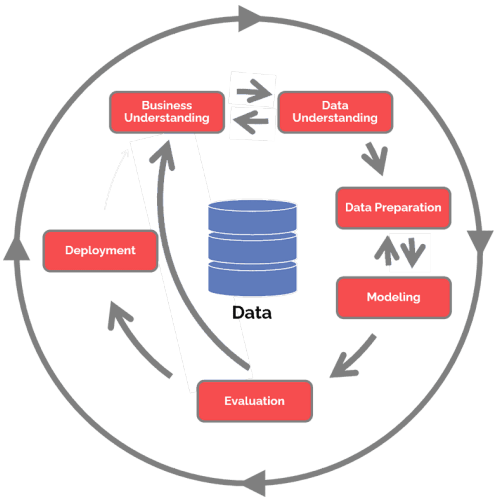

Figure 1: 6 steps in CRISP-DM, figure is obtained from datascience-pm.com.

correctly and the CAS 9 is generated on-site, the InDels of the NHEJ Repair Mechanism would effectively affect or cease the growth of Cancer cells by disrupting their duplication mechanism. However, suppose a non-primary unrelated gene has been disrupted. In that case, it might do little to no effect on the targeted cancer cell, thus wasting the cost of performing the In/Del process, sometimes even worsen the situation.

With the provided research, database on known cancer, and the datasets on the efficiency of knocking down specific genes, our goal is that: For an unknown type of cancer, provided with sufficient data, including its expression and mutation, we would try to train the most efficient Machine Learning models at predicting the on-site point, that is, the gene that should be disrupted in order to accomplish the task of disrupting the growth of said cancer cell.

2 Data Collecting and Processing

CRISP-DM is the abbreviation for the cross industry standard process for data mining, and the idea was used in our project. In general, CIRSP-DM contains 6 steps: 1. business understanding, 2. data understanding, 3. data preparation, 4. modeling, 5. evaluation and 6. depolyment, as shown in figure 1. In this work, the business understanding step is covered in the introduction section so that we're not going to reiterate it again.

The fundamental requirement of training a machine learning model correctly and accurately is to use data that verified by some agency. Here, all of cancer targeted genes were acquired by looking U.S. Food and Drug Administration (FDA) approved cancer curing drugs. Those genes were used as the genes and positive labels of training data, along side with randomly selected human genes as negative labelled genes. For the ease of us and the time limit on this project, we obtained training data which contains gene symbol and labels from GitHub (https://github.com/storm-therapeutics/CancerTargetPrediction) of the literature we followed. There are 9 cancer types in total, and features were found based on those genes defined in the training data.

Gene expressions, mutations, and essentiality are 3 primary features considered in this work based on basic biology knowledge. Gene expressions(*EB++AdjustPANCAN_ IlluminaHiSeq_RNASeqV2.geneExp.xena*) and mutations (*mc3.v0.2.8.PUBLIC.nonsilentGene.xena*) dataset were downloaded from University of California Santa Cruz (UCSC) Xena portal (https://xena.ucsc.edu). Those two datasets contain expression and mutations respectively with genes as features in column-wise and patients as observations in row-wise. Cancer types of patients were obtained from TCGA phenotype dataset (*TCGA_phenotype_denseDataOnlyDownload.tsv*) which was also downloaded from UCSC Xena portal. First, a per-cancer gene mutation rate was calculated by averaging along each genes with 1 for mutated and 0 for unmutated, and a per-cancer expression median was calculated by taking the median of each gene. Similarly, gene essentiality features were

evaluated by taking the median of each gene from gene effect dataset (*Achilles_gene_effect.csv*) downloaded from DepMap website (http://depmap.org).

In addition to those 3 primary features, there are 32 gene-gene interaction features were used in order to discovery some unknown correlation between genes and provide more information to machine learning models. To generate those 32 additional features, we first acquired all human protein-coding genes from HumanMine website (http://humanmine.org). Then, we queried the BioGRID database REST Service (https://wiki.thebiogrid.org/doku.php/biogridrest) for gene-gene interaction for each gene. By this far, we got 18,898 human progein-coding genes and 570,342 gene-gene interaction information. We then computed 32 features by applying the 32 dimensional numerical embedding of the interaction network using sequence-based embedding with diffusion graphs, finished by using Diff2Vec python script that are available on Github (https://github.com/benedekrozemberczki/diff2vec).

Multivariate feature selection was performed on 32 dimensional gene-gene interaction features in order to screen out those insignificant ones. In order to perform multivariate feature selection, the importance results of each feature were calculated using Random Forest model, and z score of each feature was calculated. Normal distributed feature importance was calculated by shuffling labels 100 times and the mean and standard deviation of this normal distribution were used for z score calculation. At the end, features whose z score is larger than 1 were used for further use, i.e., training sets.

3 Learning Methodology

3.1 Random Forest

Decision tree is an algorithm that divides data into different groups using different variables at each iteration. A Random Forest Regressor in python selects a subset of features randomly and generate decision tree to find the average in order to make predictions. While Bagging Regressor selects a subset of data randomly to generate decision trees and make predictions. In this example, however, we are allowing the default setting, thus Random Forest Regressor is using $sqrt(max\_features)$, which is 5 features, while Bagging Regressor is using all the data and all the features in one decision tree.

3.2 Logistic Regression

Logistic regression is a analysis regression that predicts the binary variables in a logistic model, by using the logistic function to calculate its coefficient and log-odd scale. As it's one of the most popular analysis regression, we also performed this method on our features.

3.3 Artificial Neural Network

An Artificial Neural Network (ANN) is a computing system that mimics animal brains. An ANN is represented as a connected directed, weighted graph that is typically organized in multiple layers. The nodes of this graph are called *neurons*. Each neuron applied an activation function to a real values input to produce a real valued output. Thus, our leaning problem translates to finding the optimal weights for the edges in the neural network. This is done by minimizing a loss function and back-propagating the gradients of this loss function.

In the context of our paper, we use a neural network designed for classification of genes into "positive" and "negative" categories.

We used a Keras based neural network with 5 fully connected layers with ReLu activation for internal layers and sigmoid activation for the classification layer. Batch Normalization and a dropout machanism was used to speed up the training process.

3.4 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is reduction technique that is used for classification problems. We use this method to do classification between classes ("positive" and "negative" categories of genes.) Also, it's used for projecting the their features from higher dimension space to lower dimension space.

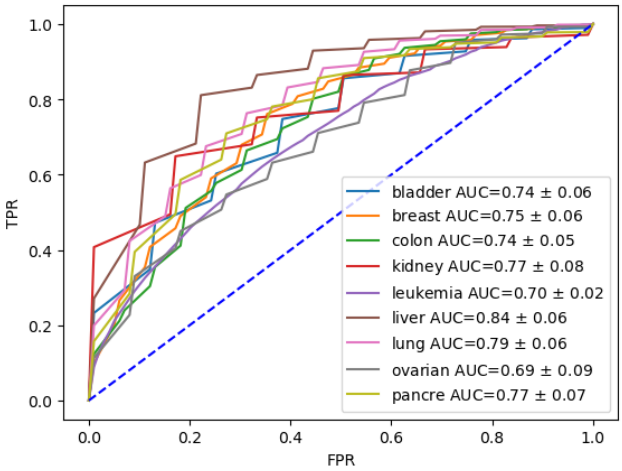

Figure 2: The ROC curve of Random Forest on 32+3 features, different Lineage with AUC score labeled.

4 Results

4.1 Random Forest

We used a Random Forest Regressor and a Bagging Regressor training on 10 to 100 different decision trees and find out that the best results comes when we choose a Random Forest Regressor with 50 decision trees. Thus we used this specs to train our data.

We split the data to 80% training and 20% testing with stratified sampling according to its label (y value) for each training iteration, and find the average for multiple iterations to construct the AUC-ROC curve. We set the number of iterations to 20 because of the limited number of data we have. A higher iteration number could cause much more repetitions. The results is shown in Figure 2

4.2 Artificial Neural Network

ANNs had differential performance for different lineages. We used the area under the ROC curve to measure performance in this paper. Performance was exceptionally high for Liver and Kidney lineages but poor for pancreatic cancer.

4.3 Logistic Regression

A logistic regression analysis is also performed on our 32 features + 3 primary features, with a random state of 42 and a split ratio of .33. The result is shown in Figure 3 below.

It does not provide the most accurate learning model, however, logistic regression do perform well on cancers that have Kidney and Liver lineages, with a Accuracy of .9 for Kidney and .88 for Liver.

4.4 Support Vector Machine

Support Vector Machine(SVM) is used with linear kernel in our work to analysis labeled gene data corresponding to each cancer type. Here the feature set including 32 embedding feature set and 3 primary features are used. Data sets for each cancer type are split by 70% and 30%. We use the ROC and AUC to study the performance of each method. The result of SVM method shows in ROC suggest a balance performance through each cancer type. There is not a single cancer getting a poor (less than 0.7) regarding to AUC.

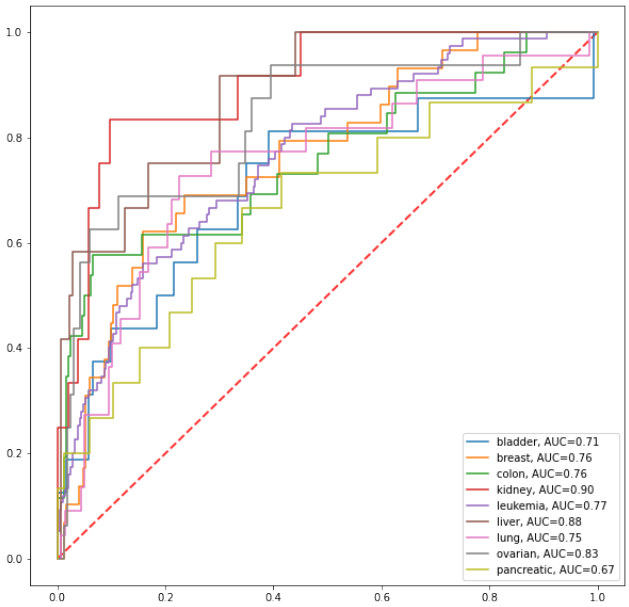

Figure 3: The ROC curve of Logistic Regression on 32+3 features, different Lineage with AUC score labeled.

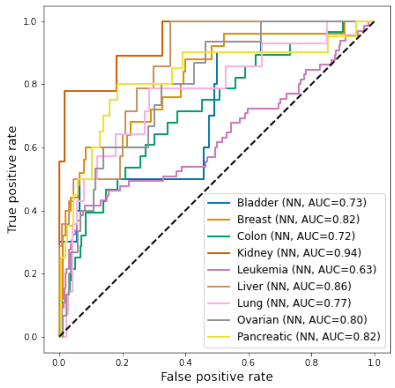

Figure 4: The ROC curve for Artificial Neural Network on 32+3 features, different Lineage with AUC score labeled.

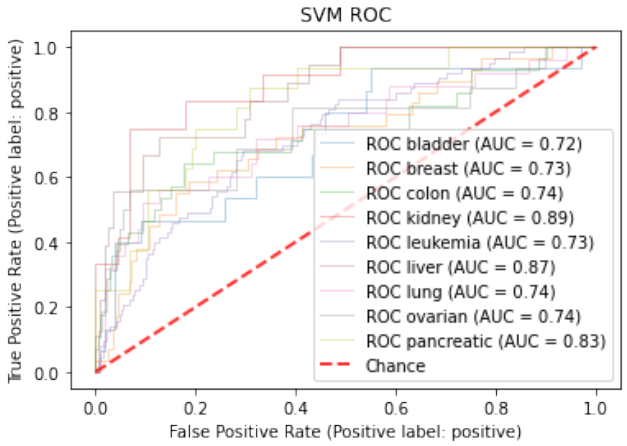

Figure 5: The ROC graph for Support Vector Machine method described in sec 4.4

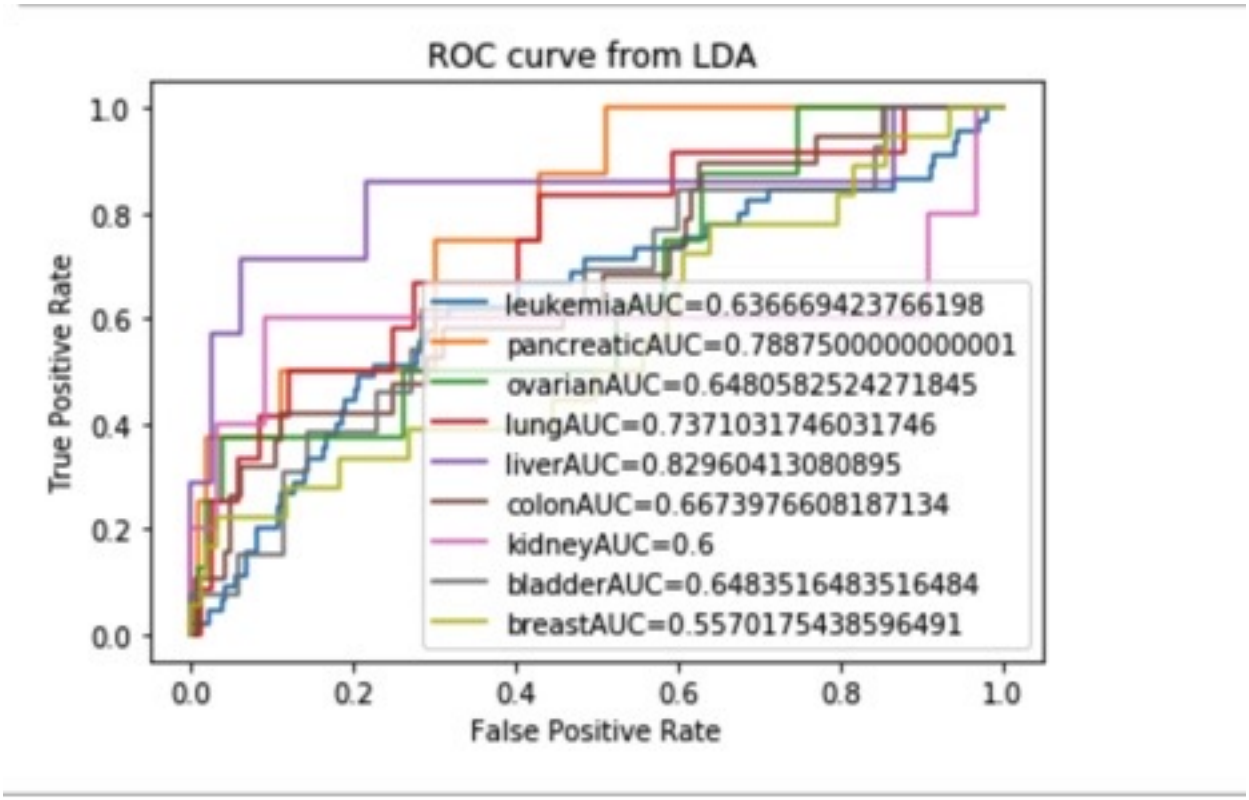

Figure 6: The ROC graph of LDA with all features

4.5 Linear Discriminant Analysis

We use LDA model to find if there is a linear combination of features to separate two classes of different organ genes. First, we use all columns of feature to be our independent variables and then try the selected feature with an importance larger than 0.75 for comparison. The results of LDA with all feature shows that it does not have a good perform with the highest AUC score 0.82.6)

The results of LDA with the selected feature shows that it has a better performance than all feature with the highest AUC score is 0.93.7

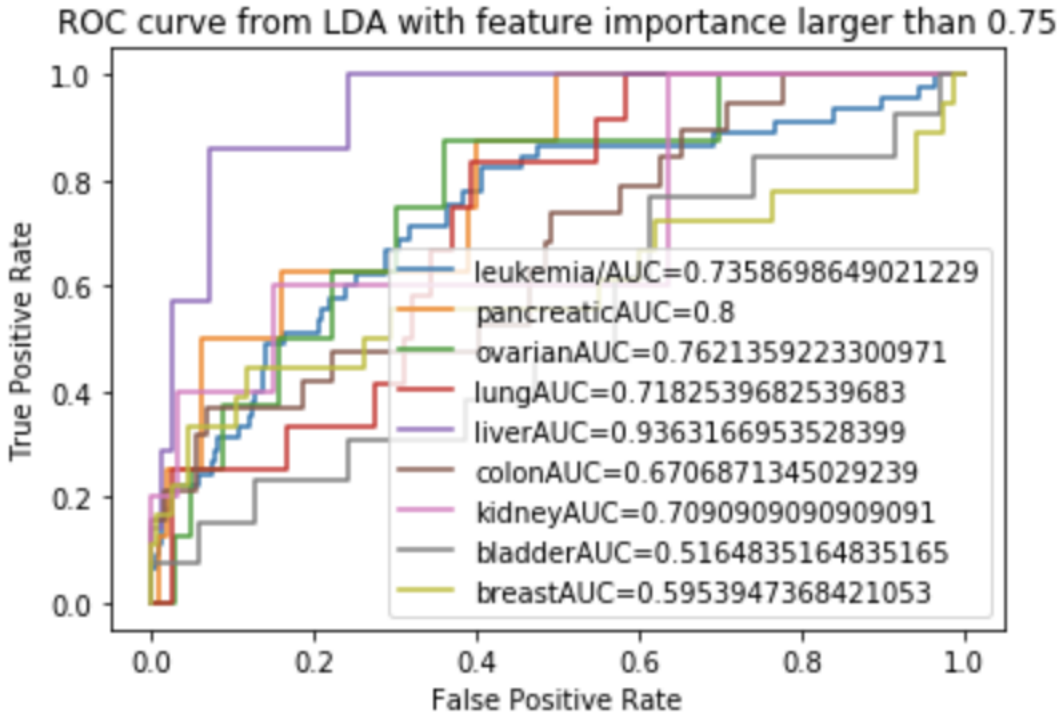

Figure 7: The ROC graph of LDA with selected features

5 Discussion and Future Work

As indicated by both our and the original paper's results, different models do not have similar performance. The source of this differential performance can be another topic for research.

Also, the data distribution was not proportional, with a 1:10 positive to negative labelled data, the algorithms tends to predict more negatives. So further work could be done either by expanding the positive labelled data or trim the negative labelled data. Compared to the original paper's drug based data set, the CRISPR KO methodology does not favor the three primary features over the network features. This suggests that CRISPR KO can potentially target cancer causing genes that are possibly overlooked by the three primary features. This has massive impact in classification of cancer patients into different treatment groups.

On the other hand, as indicated by the previous paper, addition of network based features can add bias towards genes that are well studied because the construction of network based features is highly dependent on the choice of gene-interaction network.

References

[1] Bazaga, A. & Leggate, D. & Weisser, H. (2020) Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology. *Sci Rep* 10, 10787 (2020). https://doi.org/10.1038/s41598-020-67846-1