MATH 4570 - Matrix Methods in Data Analysis and Machine Learning
December 7, 2022

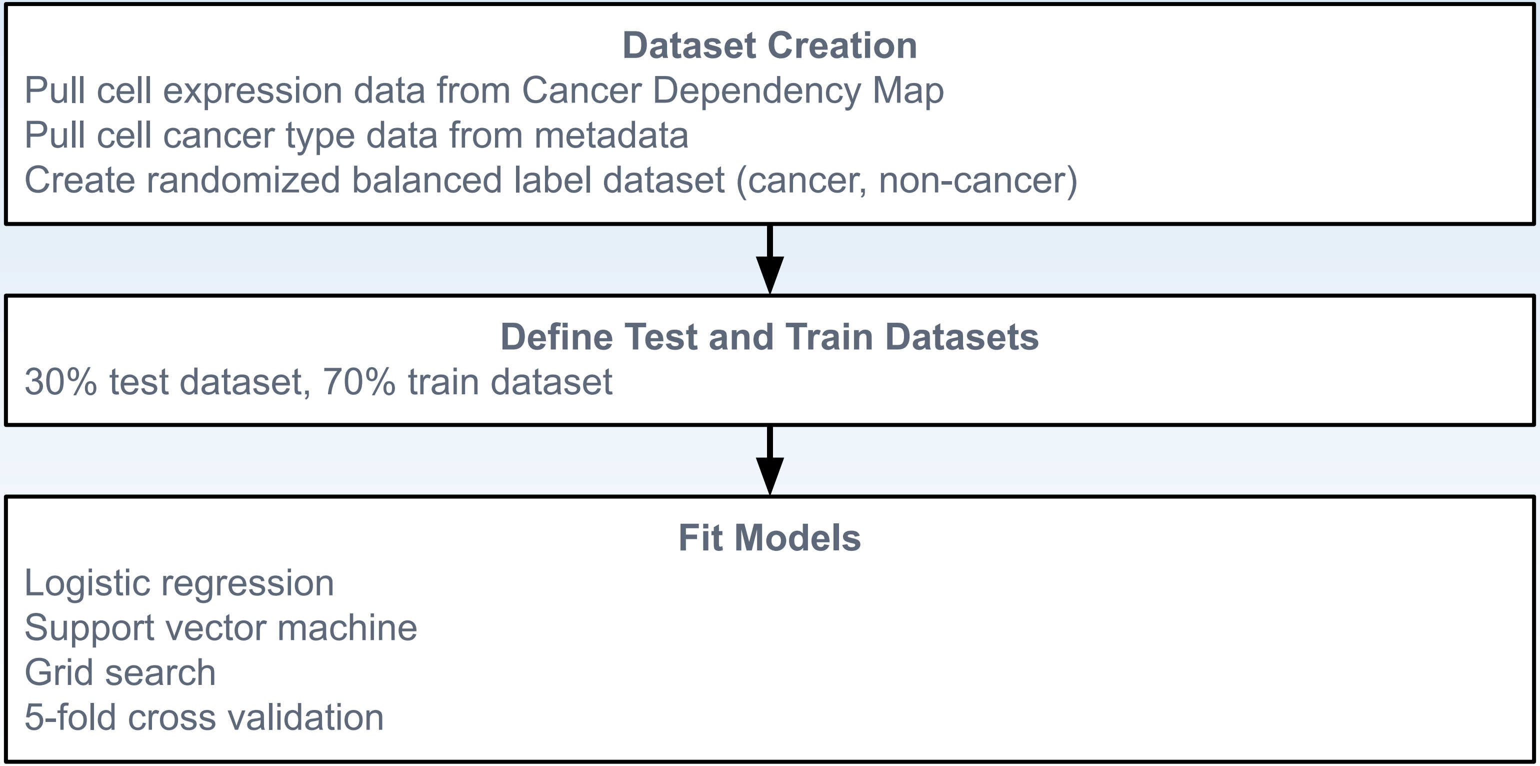
Machine Learning - Finding Novel Intrinsic Oncology Targets in Biology

Northeastern University
Rien Looijenga, Mark Larson, Chinmai Deo

Introduction

- Cancer
  - Disease associated with uncontrolled growth of cells
  - Has potential to destroy normal body tissue
  - Often triggered by genetic mutations
- Development of cancer therapies
  - Cancer drugs act on cancer-causing genes within cells
  - **Problem:** Identifying cancer-causing genes is expensive and time consuming
- **Goal 1:** Mine genetic datasets with machine learning methods to predict whether a cell is cancerous
- **Goal 2:** Use machine learning methods to mine genetic datasets for cancer vulnerabilities by classifying target and non-target genes.
- **Goal 3:** Mine genetic datasets with machine learning methods to predict cancer type

Part I Methodology: Classifying Cancerous Cells



Dataset Creation

Pull cell expression data from Cancer Dependency Map
Pull cell cancer type data from metadata
Create randomized balanced label dataset (cancer, non-cancer)

Define Test and Train Datasets

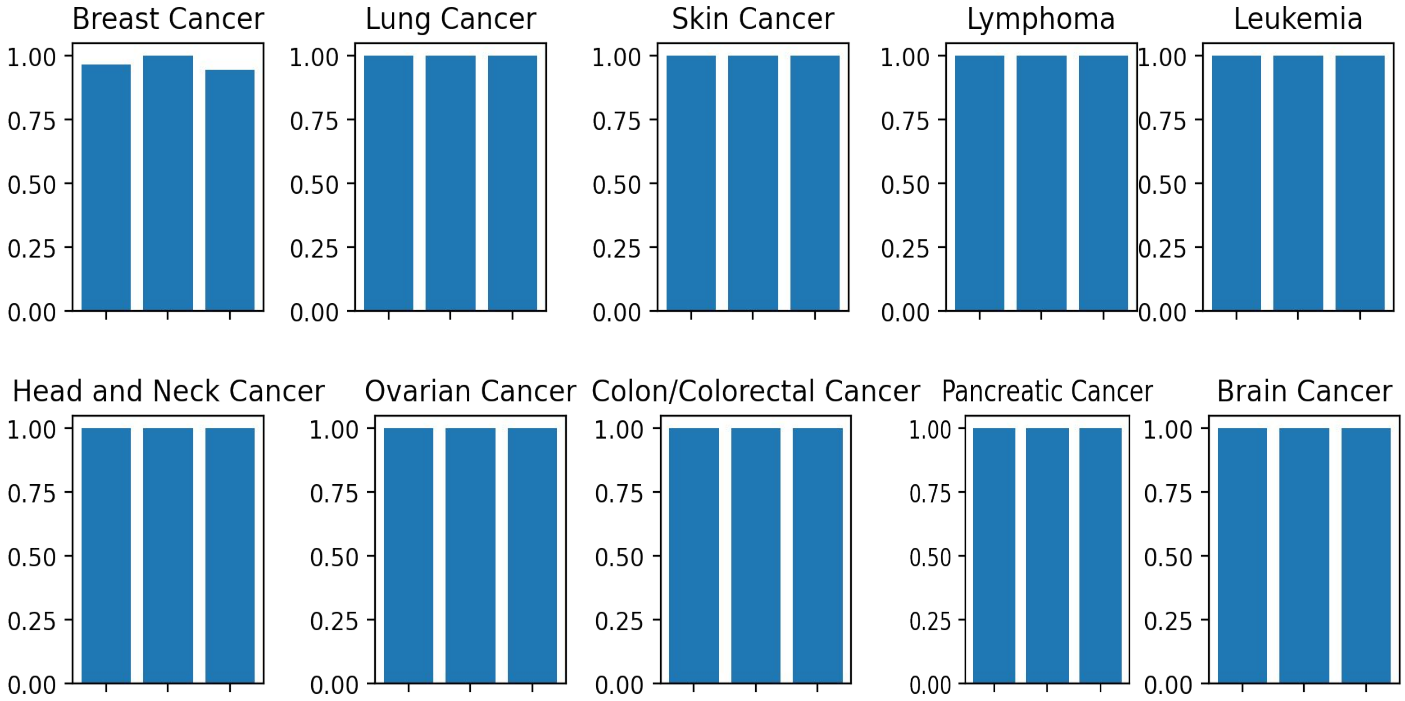
30% test dataset, 70% train dataset

Fit Models

Logistic regression
Support vector machine
Grid search
5-fold cross validation

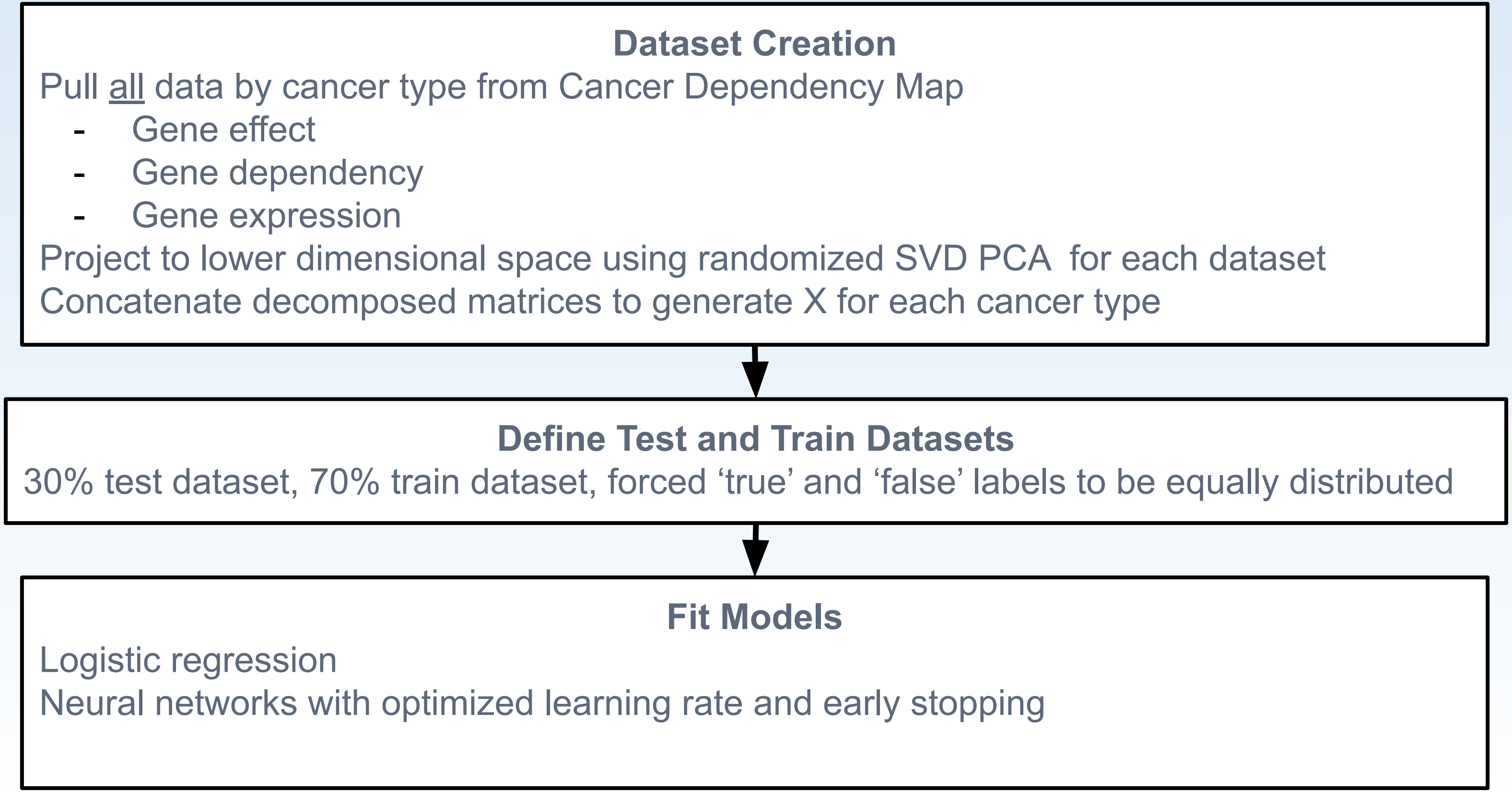
Part I: SVM Results

Classifying Cancerous Cells



Accuracy, precision, and recall for SVM. Similarly perfect results from LR

Part II Methodology: Classifying Oncology Target Genes



Dataset Creation

Pull all data by cancer type from Cancer Dependency Map

- Gene effect
- Gene dependency
- Gene expression

Project to lower dimensional space using randomized SVD PCA for each dataset

Concatenate decomposed matrices to generate X for each cancer type

Define Test and Train Datasets

30% test dataset, 70% train dataset, forced ‘true’ and ‘false’ labels to be equally distributed

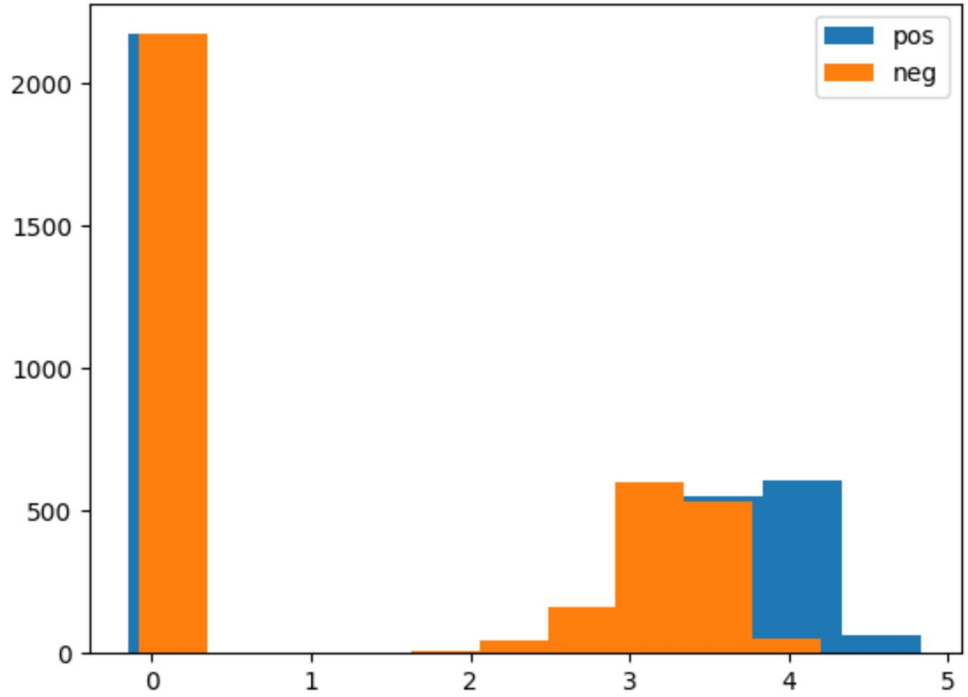
Fit Models

Logistic regression

Neural networks with optimized learning rate and early stopping

Part II: Label Data

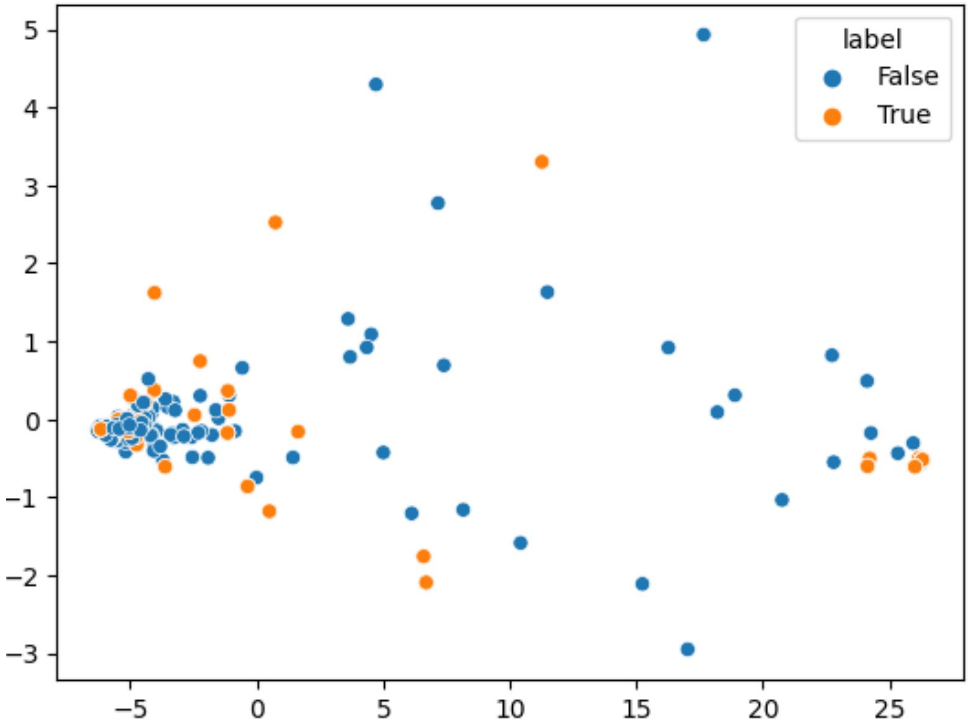
Classifying Oncology Target Genes



Differences in cells by target gene label

Part II: Label Data

Classifying Oncology Target Genes



Concatenated PCA data transform (Bladder Cancer Genes)

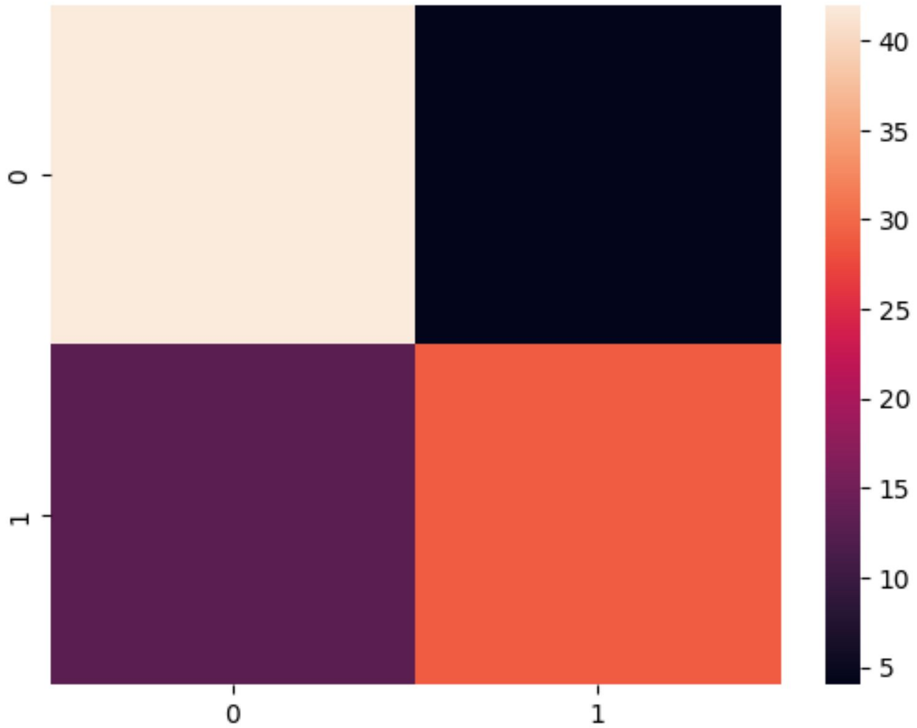
Part II: Model Comparison

Classifying Oncology Target Genes

Cancer Type	Logistic Accuracy	Neural Network Accuracy	Difference
All	0.5743	0.6658	0.0915
Bladder	0.8068	0.8864	0.0796
Breast	0.7211	0.7551	0.034
Colon	0.6692	0.6842	0.015
Kidney	0.7826	0.8261	0.0435
Leukemia	0.6008	0.6532	0.0524
Lung	0.6569	0.6275	-0.0294
Ovarian	0.7129	0.802	0.0891
Pancreatic	0.7355	0.8182	0.0827
Liver	0.62	0.608	-0.012
Average	*0.68801*	*0.73265*	*0.04464*

Part II: Logistic Regression

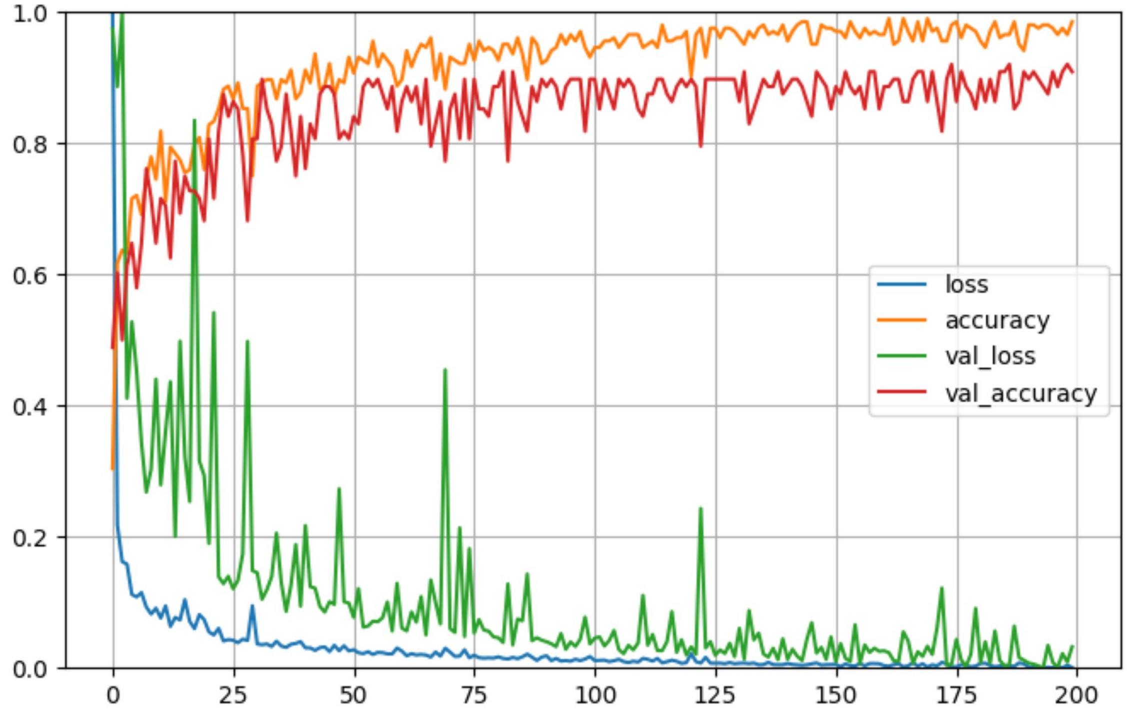
Classifying Oncology Target Genes



Logistic regression confusion matrix

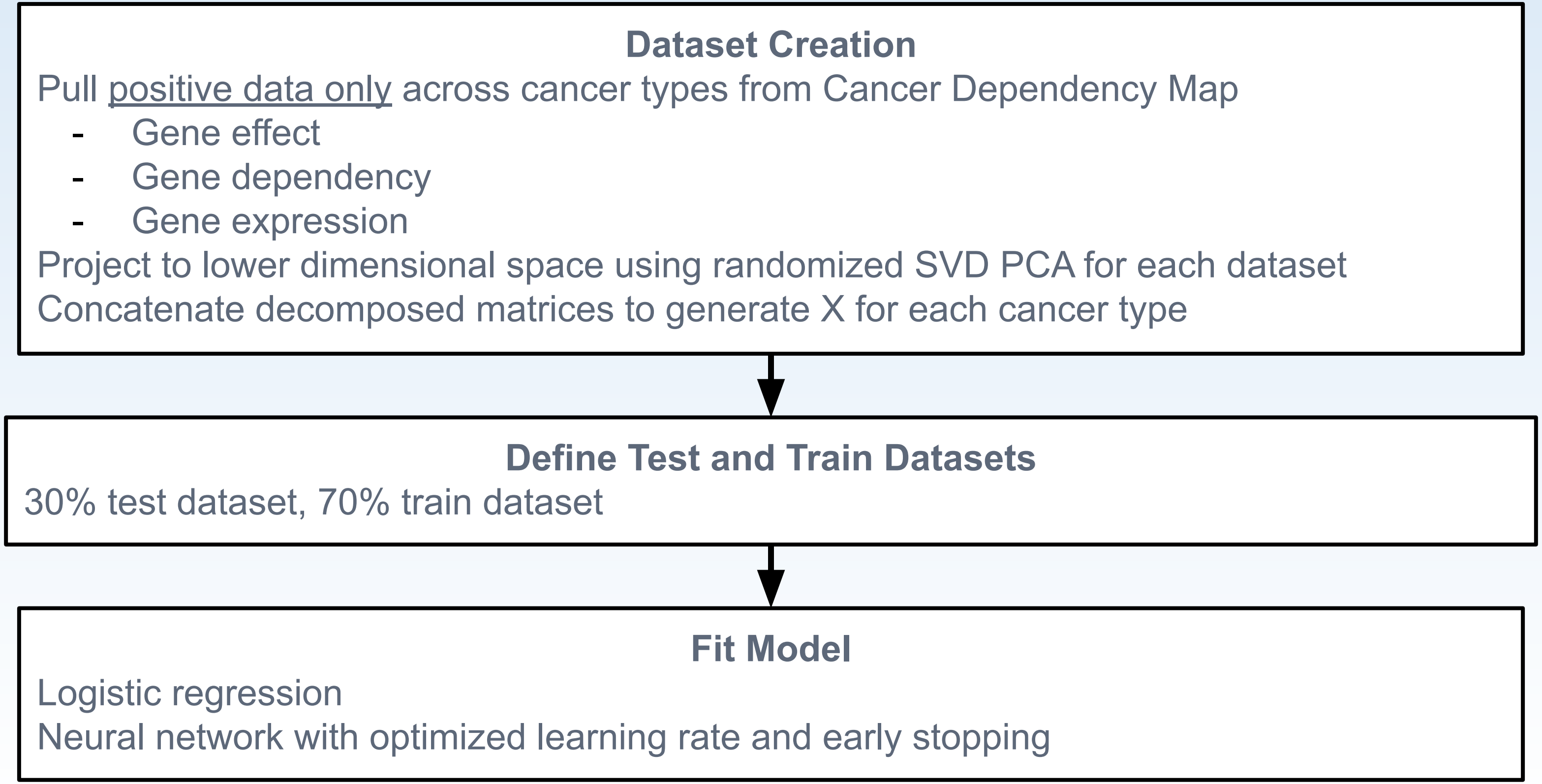
Part II: Neural Network

Classifying Oncology Target Genes using Categorical Cross Entropy Loss / SGD optimizer



Neural Network Loss Performance

Part III Methodology: Predicting Cancer Type



Dataset Creation

Pull positive data only across cancer types from Cancer Dependency Map

- Gene effect
- Gene dependency
- Gene expression

Project to lower dimensional space using randomized SVD PCA for each dataset

Concatenate decomposed matrices to generate X for each cancer type

Define Test and Train Datasets

30% test dataset, 70% train dataset

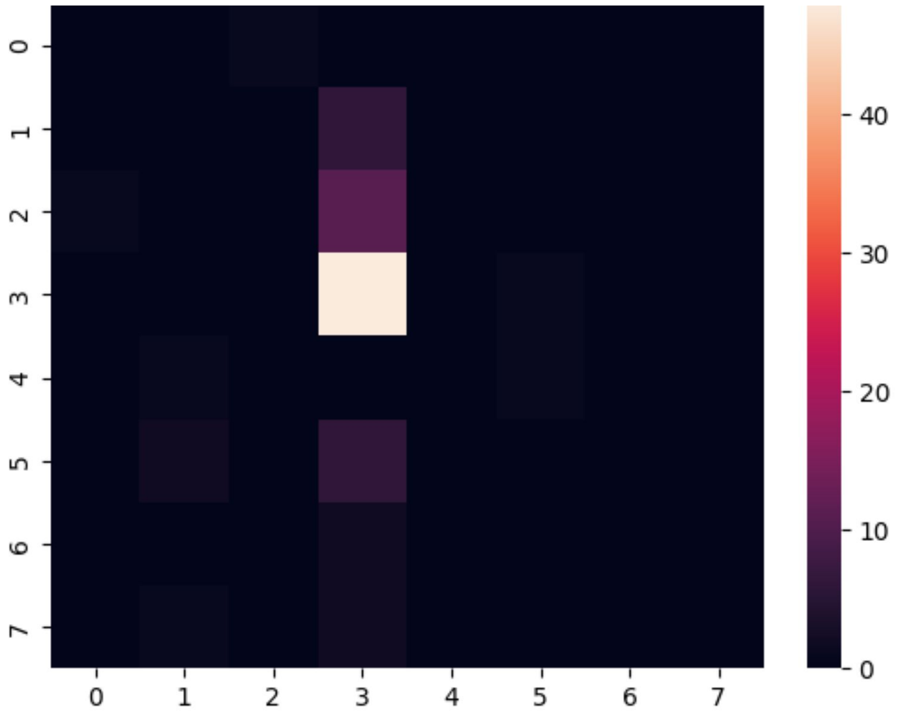
Fit Model

Logistic regression

Neural network with optimized learning rate and early stopping

Part III: Logistic Regression

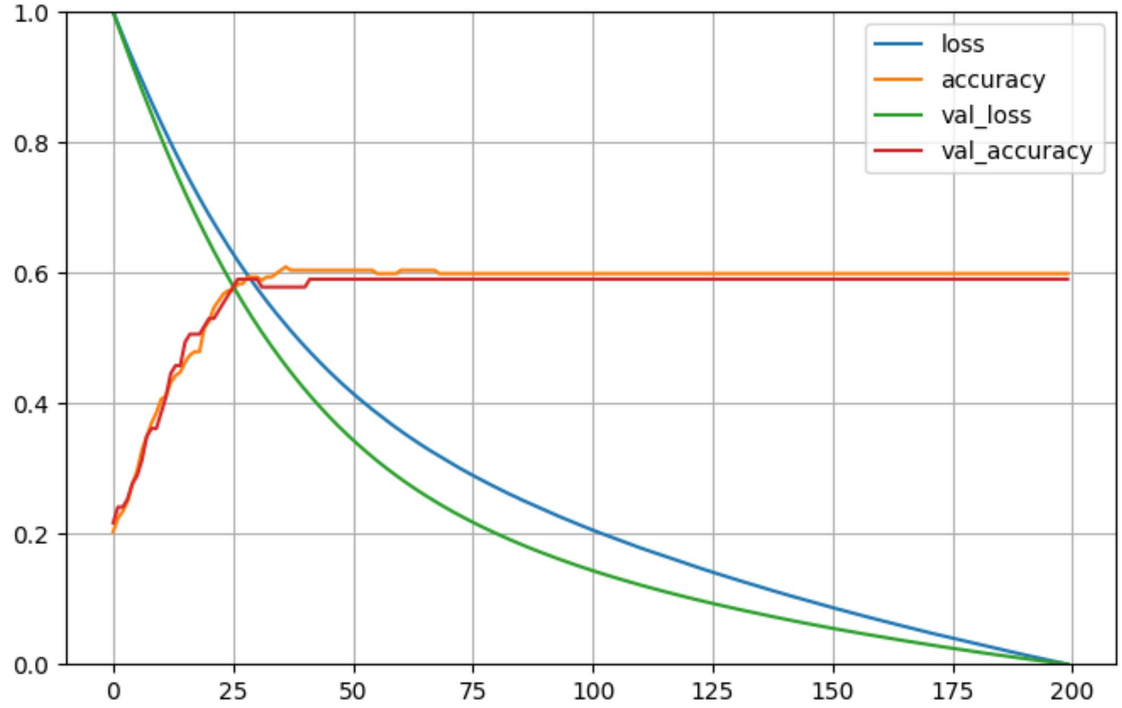
Predicting Cancer Type



Logistic regression confusion matrix

Part III: Neural Network

Predicting Cancer Type using Categorical Cross Entropy Loss / SGD optimizer



Neural Network Loss Performance | Accuracy = 0.5974

MATH 4570 - Matrix Methods in Data Analysis and Machine Learning
December 7, 2022

Machine Learning - Finding Novel Intrinsic Oncology Targets and Biology

Northeastern University
Rien Looijenga, Mark Larson, Chinmai Deo