

Machine Learning for Cancer Diagnosis on Microarray Data

Ty Pacchione, Danny Rollo, Alexander Vandenberg,
and Mitchell Wadas

December 7, 2023

Background

- Microarrays measure the expression of thousands of genes simultaneously.
- Microarray data is increasingly being used for early diagnosis of cancer.
- Researchers are struggling to extract meaningful information from so much data.
- We propose an application of machine learning to make diagnoses from microarray data.

Data

- Raw data was sourced from the Structural Bioinformatics and Computational Biology Lab's CuMiDa database [?]. Each file corresponds to a specific type of cancer and set of genes.

samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at
306	adenoma	9.4431086	3.65359337	5.08776789	7.61927763	3.78670578
307	adenoma	9.34227295	3.71458533	5.44399744	7.10476575	3.80912899
308	adenoma	9.1484732	3.69324912	5.17350663	7.53621486	3.7143997

- We downloaded approximately 40 datasets corresponding to different cancer / gene set combinations.
- We combined them into two datasets, one corresponding to each gene set.

Procedure

- We formulated the diagnosis as a multi-class classification problem.
- A notable characteristic of our data is the number of features, between 30,000 and 50,000.
- We compared traditional ML (logistic regression with PCA) and deep learning.
- We compared different set of genes (as features) to determine which has more diagnostic power.

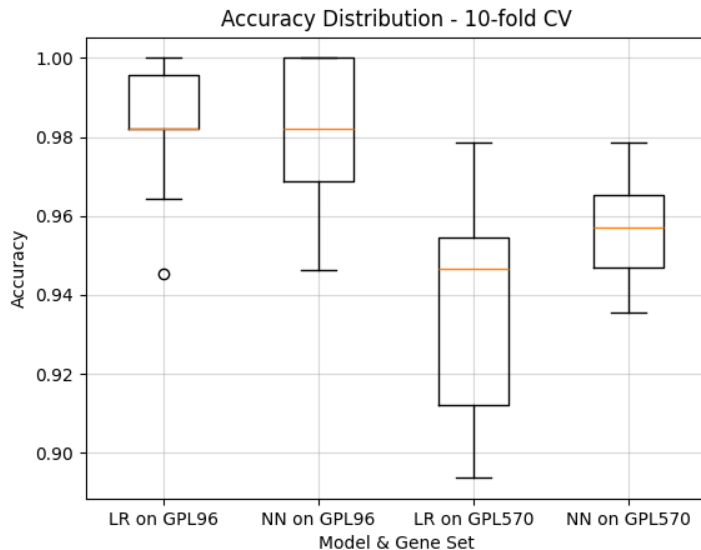
Methods - Logistic Regression with PCA

- We decomposed the data into 50 principal components which captured roughly 95% of the variance (efficient reduction).
- We applied logistic regression to the principal components.

Methods - Neural Network

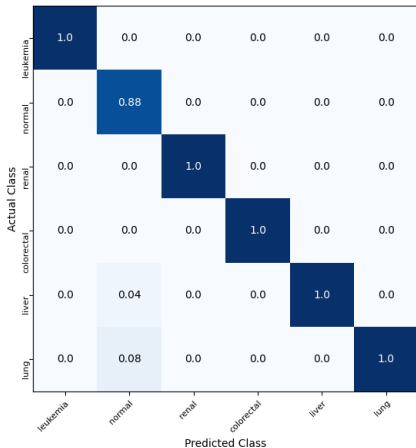
- Instead of using PCA, we gave the network the full feature set, allowing it to perform feature extraction.
- We performed grid search hyperparameter optimization to find the optimal neural architecture.

Cross Validation Results

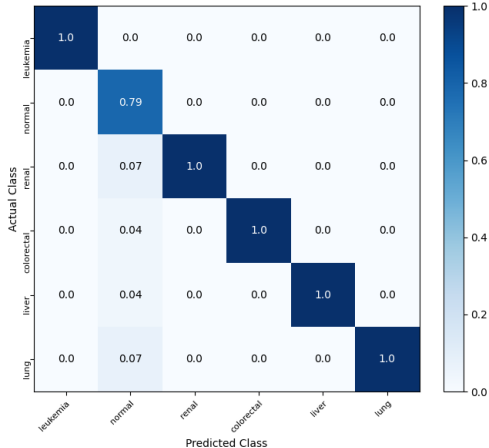


Confusion Matrix - GPL96

LR on GPL96 - Confusion Matrix

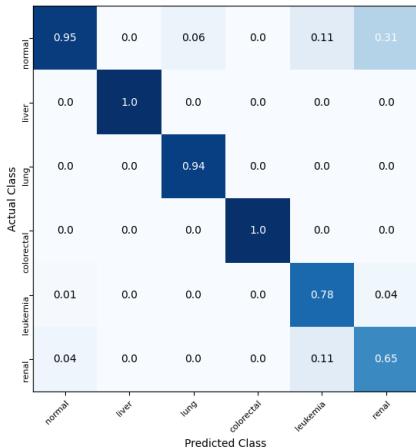


NN on GPL96 - Confusion Matrix

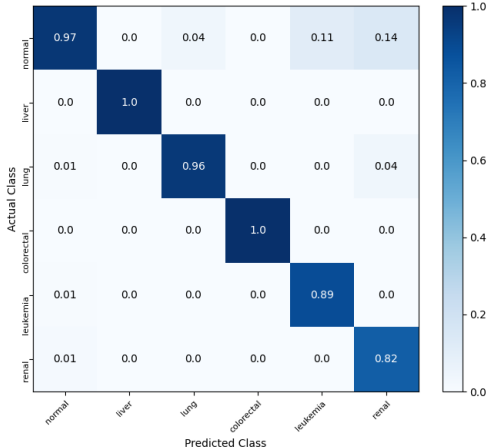


Confusion Matrix - GPL570

LR on GPL570 - Confusion Matrix

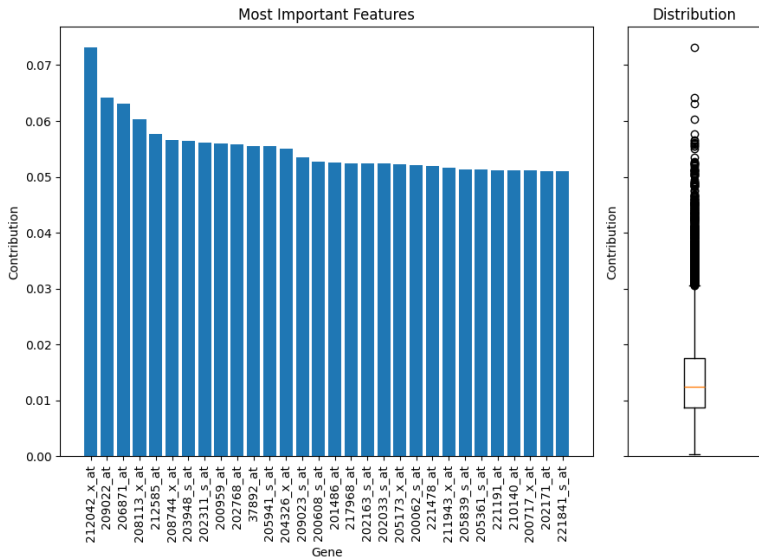


NN on GPL570 - Confusion Matrix

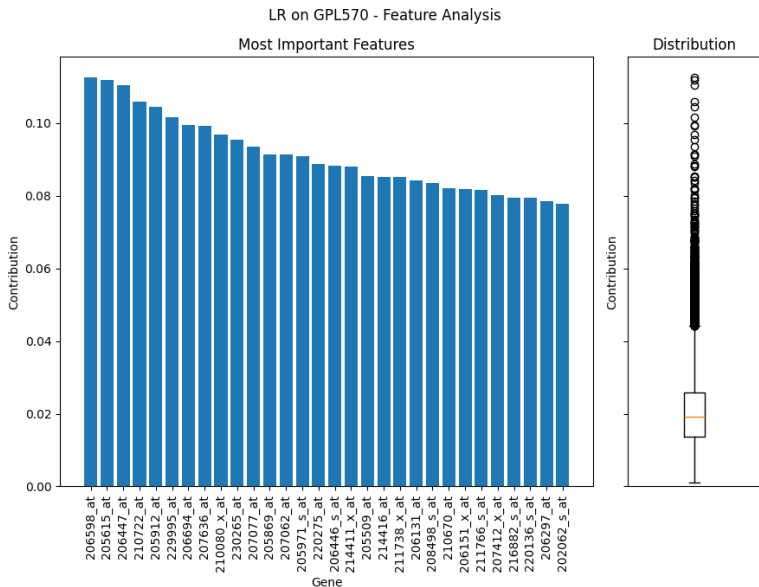


Feature Analysis - GPL96

LR on GPL96 - Feature Analysis



Feature Analysis - GPL570



References