Machine Learning for Cancer Diagnosis on Microarray Data

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

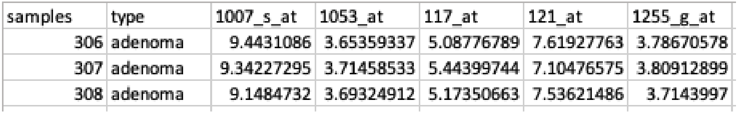
December 7, 2023

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 [1]. 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.

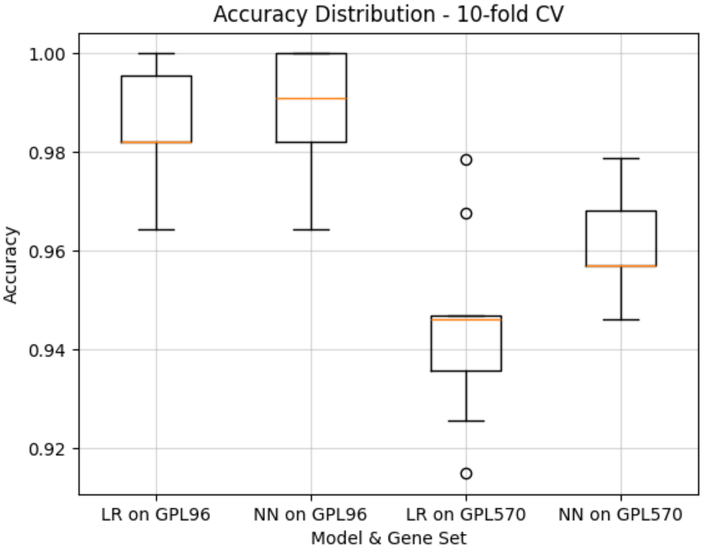
Methods - Logistic Regression with PCA

- 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.
- This resulted in an accurate diagnosis in approximately 95% of cases.

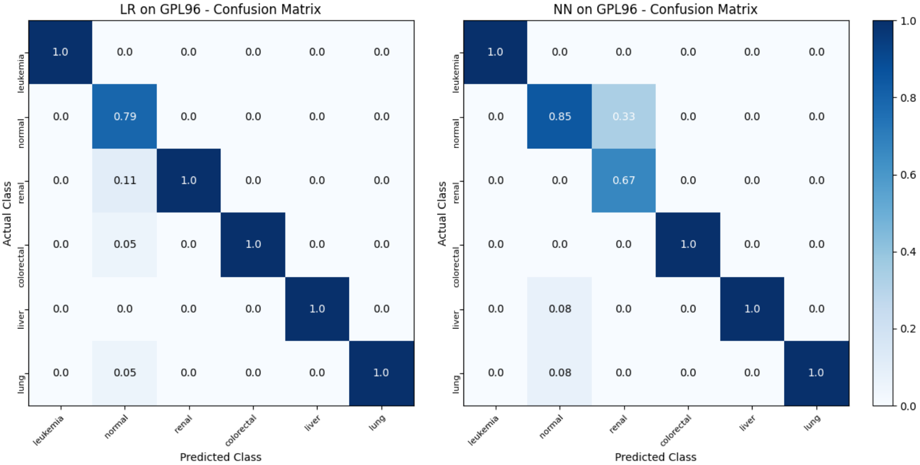
Methods - Neural Network

- 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.
- The optimal architecture was determined to be 5 hidden layers with 50 neurons in each with ReLU activation between each two layers.
- With this architecture, we achieved accurate diagnosis in approximately 97% of cases.

Cross Validation Results



Confusion Matrix - GPL96



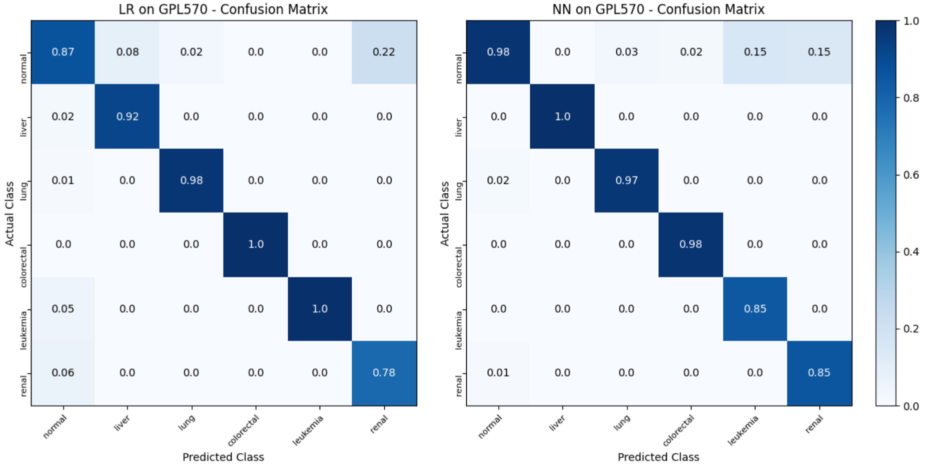
LR on GPL96 - Confusion Matrix

	leukemia	normal	renal	colorectal	liver	lung
leukemia	1.0	0.0	0.0	0.0	0.0	0.0
normal	0.0	0.79	0.0	0.0	0.0	0.0
renal	0.0	0.11	1.0	0.0	0.0	0.0
colorectal	0.0	0.05	0.0	1.0	0.0	0.0
liver	0.0	0.0	0.0	0.0	1.0	0.0
lung	0.0	0.05	0.0	0.0	0.0	1.0

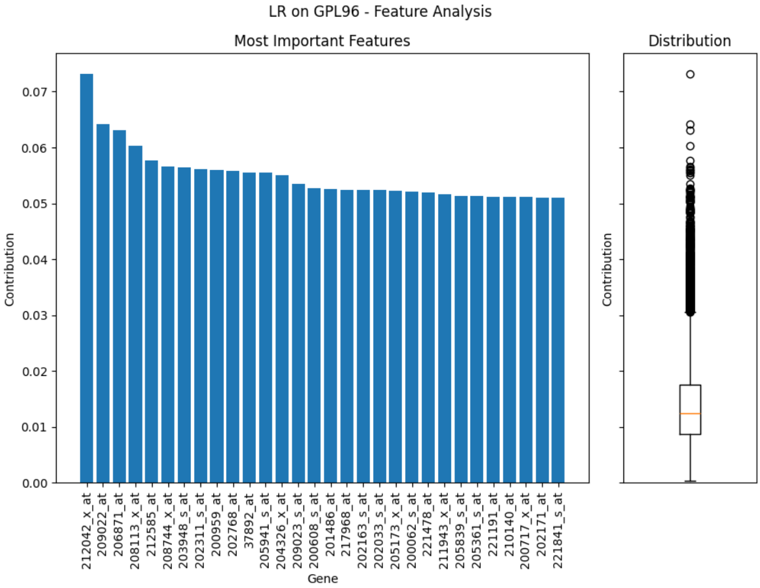
NN on GPL96 - Confusion Matrix

	leukemia	normal	renal	colorectal	liver	lung
leukemia	1.0	0.0	0.0	0.0	0.0	0.0
normal	0.0	0.85	0.33	0.0	0.0	0.0
renal	0.0	0.0	0.67	0.0	0.0	0.0
colorectal	0.0	0.0	0.0	1.0	0.0	0.0
liver	0.0	0.08	0.0	0.0	1.0	0.0
lung	0.0	0.08	0.0	0.0	0.0	1.0

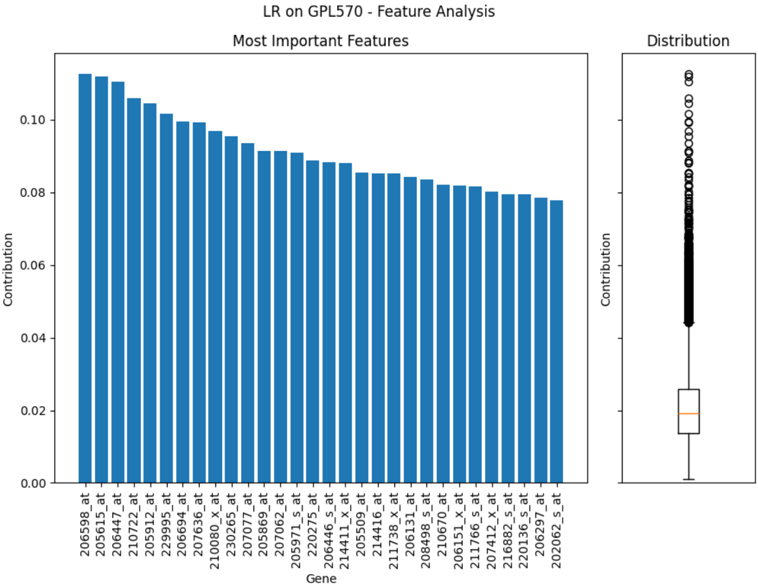
Confusion Matrix - GPL570



Feature Analysis - GPL96



Feature Analysis - GPL570



Conclusions

- We found that Neural Networks generally outperformed traditional machine learning techniques on this high dimensional multi-class problem.
- By comparing performance across datasets, we found that the GPL96 gene set was more suitable for cancer diagnosis.
- We observed that in both datasets, with both models, it was particularly hard to diagnose renal (kidney) cancer.

References

[1] Bruno César Feltes et al. “CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research”. In: *Journal of Computational Biology* 26.4 (2019). PMID: 30789283, pp. 376–386. DOI: 10.1089/cmb.2018.0238. eprint: https://doi.org/10.1089/cmb.2018.0238. URL: https://doi.org/10.1089/cmb.2018.0238.

[2] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.