Machine Learning Based Classification of Cancer Types and Identification of Gene Biomarkers from RNA-Seq Data

Lucas Bichescu

Dartmouth Health

ABSTRACT

- Gene expression profiling of thousands of genes, enabling precise distinction between five cancer types
- Logistic Regression and SVM proved to be very effective (>99% accuracy)
- Statistical methods (ANOVA, Recursive Feature Elimination) identify highly discriminative genes.
- Visualizations (heatmaps, volcano plots) highlight expression patterns and molecular signatures.

INTRODUCTION

Understanding Cancer:

- Being the second leading cause of death, cancer accounts for over 10 million deaths per year
- The WHO Global Cancer Report projects a 57 % of global cancer incidence over the next two decades
- relies on visual examination of tissue structure and cell morphology, which can be subjective and may not capture the full molecular heterogeneity of tumors
- By using thousands of gene profiles, RNA sequencing enables complex and highly informative identification of molecular subtypes that may not be apparent through histological analysis

RNA-Seq:

- RNA is first isolated from a sample and Poly-A selection is used to capture mRNAs. RNA is then converted into complementary DNA (cDNA) using reverse transcriptase such that sequencing machines can interpret the data
- Conversion of RNA to DNA allow for easy analysis of transcription errors, alternative splicing, RNA editing, and more that DNA information alone would not display

Data:

- Included in the UCI Learning Repository made first by the TCGA Pan Cancer analysis project
- Contains five cancer types: (BRCA) Breast Invasive Carcinoma, (KIRC) Kidney Renal Clear Cell Carcinoma, (COAD) Colon Adenocarcinoma, (LUAD) Lung Adenocarcinoma, (PRAD) Prostate Adenocarcinoma
- 801 instances, 20,531 dummy gene expression features (gene_1,gene_2, etc.)

METHODS

Goal:
Use RNA-seq data and machine learning to distinguish between five different types of cancer, using a number of advanced data analysis approaches to gain insight on underlying genetic biomarkers

Preprocessing:

- Data is log transformed to stabilize variance having more of an impact on larger values allowing for more fair and balanced data
- Z-score is then applied, standardizing gene expression values across samples making them directly comparable

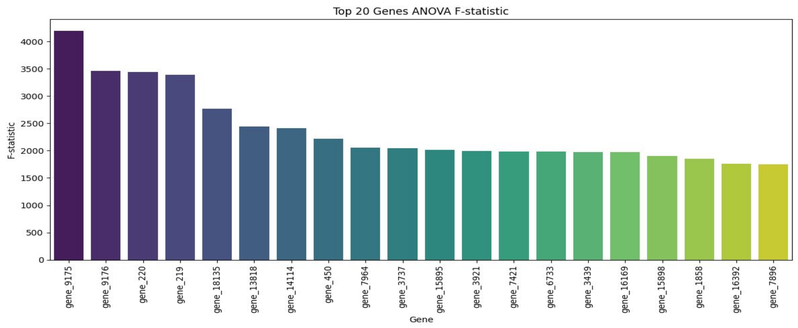
K-fold CV:

- Splits dataset into k folds, k-1 partitioned to training, one for testing; repeated for 5 folds such that every sample is used once as test data
- Used for easy model performance estimation, training, and hyperparameter tuning with grid search

Gene Significance Plots:

- I use a combination of descriptive methods such as ANOVA, heatmaps, volcano plots, and Recursive Feature Elimination

RESULTS



ANOVA is used to identify genes whose expression differs significantly across multiple cancer types. Genes such as 9175, 9176, and 220 show the highest f-stat values, indicating those of highest importance

Figure 1: ANOVA F-stat

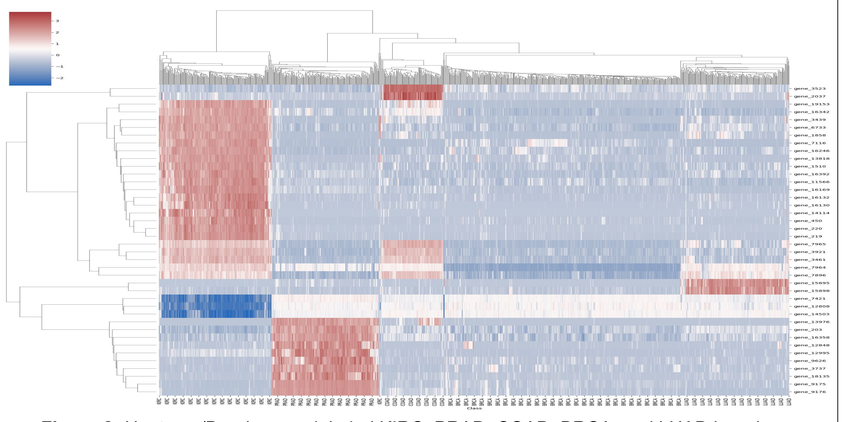


Figure 2: Heatmap/Dendrogram labeled KIRC, PRAD, COAD, BRCA, and LUAD in order on the x-axis with genes on the y-axis

Using genes selected from ANOVA f-stat, clusters of relevant genes can be discerned through darker shades of red,For example, some of the most relevant genes of KIRC are genes 1510, 13818,and 16392 with lesser important genes 3461 and 7964 sharing similar relevancy to COAD

Model:	Accuracy:	F1 Score:	Recall:
Logistic Regression (lbfgs solver)	0.9975	0.9961	0.9975
LinearSVC	0.9988	0.9987	0.9988
Decision Tree	0.9775	0.9771	0.9705
KNN	0.9950	0.9950	0.9963

Figure 4: Prediction Accuracy Model Comparisons

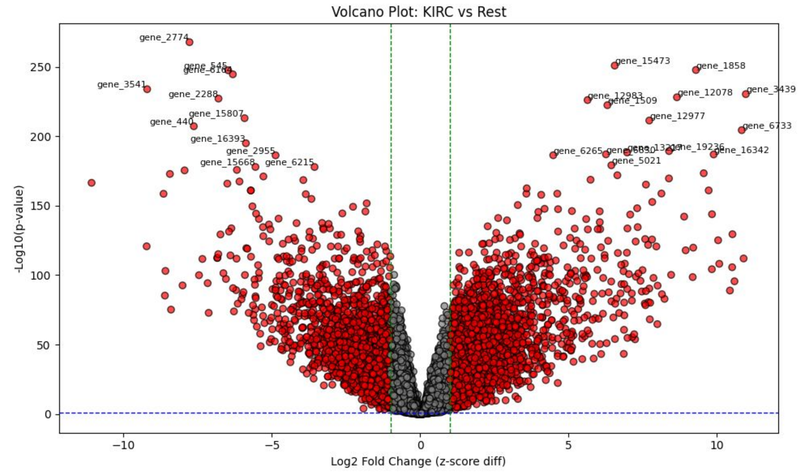
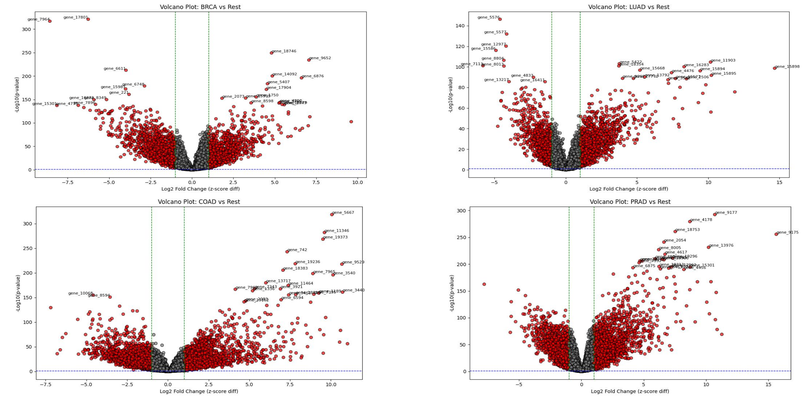


Figure 5: Volcano plot demonstrating genes important to a cancer type (Right) vs. genes important to remainder cancer types (Left)
Plots: BRCA (Top-Left), COAD (Bottom-Left), LUAD (Top-Right), PRAD (Bottom-Right)



DISCUSSION

- Through in depth analysis of the plots, patterns of genes begin to show up on every plot. For example:

KIRC: gene_3439, gene_1858, gene_16358

- Volcano plots which are individually tailored to the cancer type reveal even more information on gene biomarkers
- All models exhibit exceptional performance with Logistic Regression and Linear SVC demonstrating almost near impeccable accuracy

FUTURE WORK

- Experiment with normal tissue and other cancer types with proper gene titles
- Build a web app to easily display data analysis plots and directly communicate information to user

CONCLUSION

- By demonstrating machine learning capabilities in prediction and plotting of RNA-seq data, a foundation can be established that will facilitate automated genetic biomarker information

Special Thanks to Edit, Marietta, and Dr. Joshua Levy

References

edit.