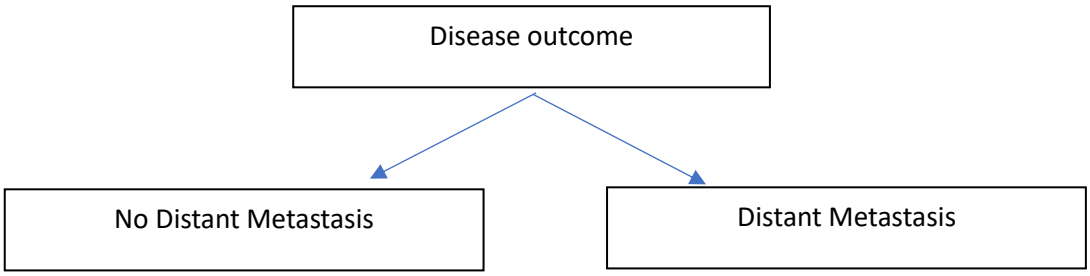
Final Progress Report (27 April 2024)

Reference Paper:

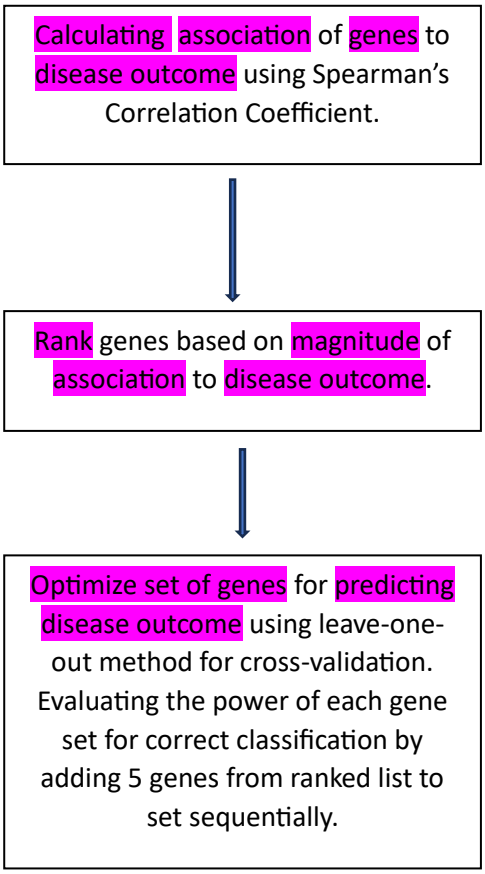
Gene expression profiling predicts clinical outcome of breast cancer.

Objective:

Finding biomarkers that predict disease outcome i.e distant metastasis in Kidney Renal Clear Cell Carcinoma (cBioPortal).



3-Step Methodology for Supervised Classification:



Data Sets: Kidney

Data Set	Microarray (M0 group)	Microarray (M1 group)	RNASeq (M0 group)	RNASeq (M1 group)	Link
Kiney Renal Clear Cell Carcinoma (TCGA, Firehose Legacy)	67	5	422	79	https://www.cbioportal.org/study/clinicalData?id=kirc_tcga
Kidney Renal Clear Cell Carcinoma (TCGA, PanCancer Atlas)	NA	NA	401	78	https://www.cbioportal.org/study/clinicalData?id=kirc_tcga_pan_can_atlas_2018

Table 1: Data Sets Information

Results:

Firehose Legacy Dataset			**PanCancer Atlas Dataset**		
Correlation	**Genes**	**Functional Annotations**	**Genes**	**Correlation**	**Functional Annotations**
0.260256	OR4A47	Olfactory Receptor Family 4 Subfamily A Member 47 (GeneCards)	0.283016	CLDN22	Claudin 22, integral membrane proteins and components of tight junction strands (GeneCards)
0.252586	OR51T1	Olfactory Receptor Family 51 Subfamily T Member 1 (GeneCards)	0.27459	DCD	Dermcidin, The C-terminal peptide is expressed in sweat and has antibacterial properties (GeneCards
0.250252	OR51S1	Olfactory Receptor Family 51 Subfamily S Member 1 (GeneCards)	0.24677	ADAM3A	ADAM Metallopeptidase Domain 3A (Pseudogene) (GeneCards)
0.229014	FEZF1	FEZ Family Zinc Finger 1 (GeneCards)	0.22896	RQCD1	CCR4-NOT Transcription Complex Subunit 9 (GeneCards)
0.222449	METTL21C	Methyltransferase 21C, AARS1 Lysine (GeneCards)	0.228029	AKAP10	A-Kinase Anchoring Protein 10 (GeneCards)
0.21819	ATP4B	ATPase H+/K+ Transporting Subunit Beta (GeneCards)	0.218204	SMC3	Structural Maintenance Of Chromosomes 3 (GeneCards)
0.189936	CYP4F30P	Cytochrome P450 Family 4 Subfamily F Member 30, Pseudogene (GeneCards)	0.211493	RAB11FIP2	R AB11 Family Interacting Protein 2 (GeneCards)
0.189933	PATE1	Prostate And Testis Expressed 1 (GeneCards)	0.209383	HIAT1	Histone Acetyltransferase 1 (GeneCards)
0.186319	ZFX	Zinc Finger Protein X-Linked (GeneCards)	0.208705	KRTAP5-5	Keratin Associated Protein 5-5 (GeneCards)
0.182832	CSNK1A1L	Casein Kinase 1 Alpha 1 Like (GeneCards)	0.203884	RPL28	Ribosomal Protein L28 (GeneCards)

Table 2: Correlation and Gene Information

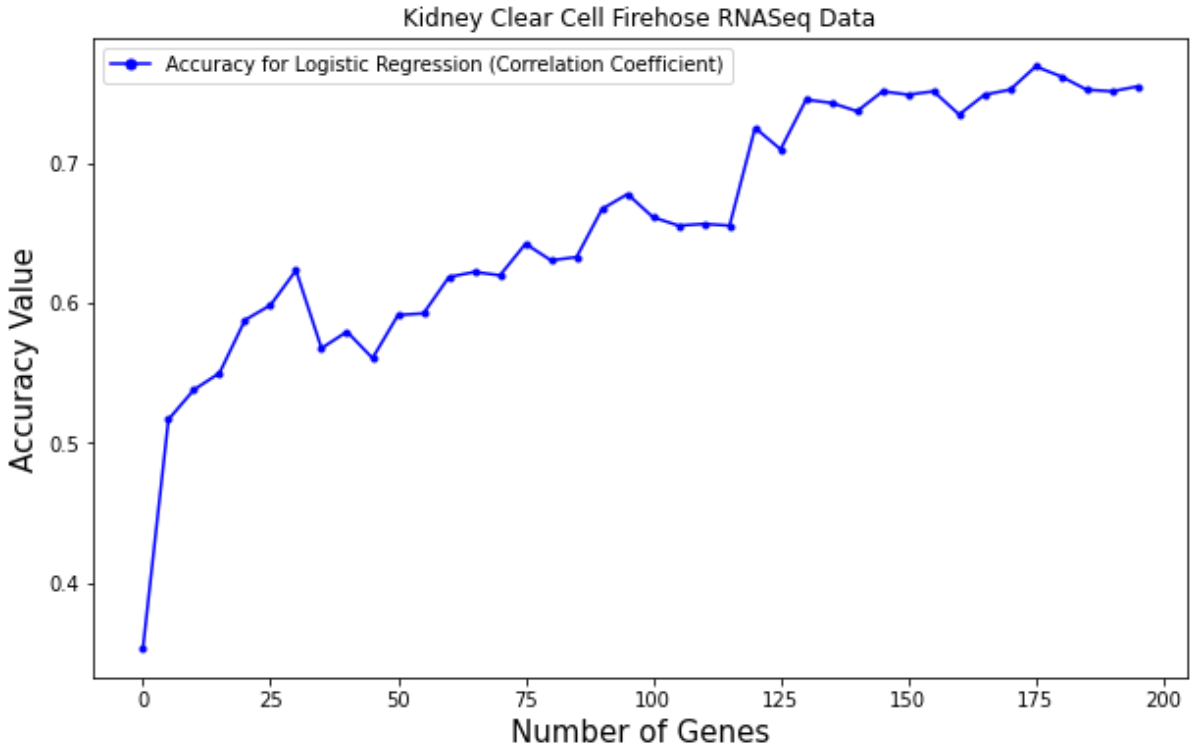


Figure 1: Accuracy against Number of Genes for TCGA Firehose Legacy Dataset

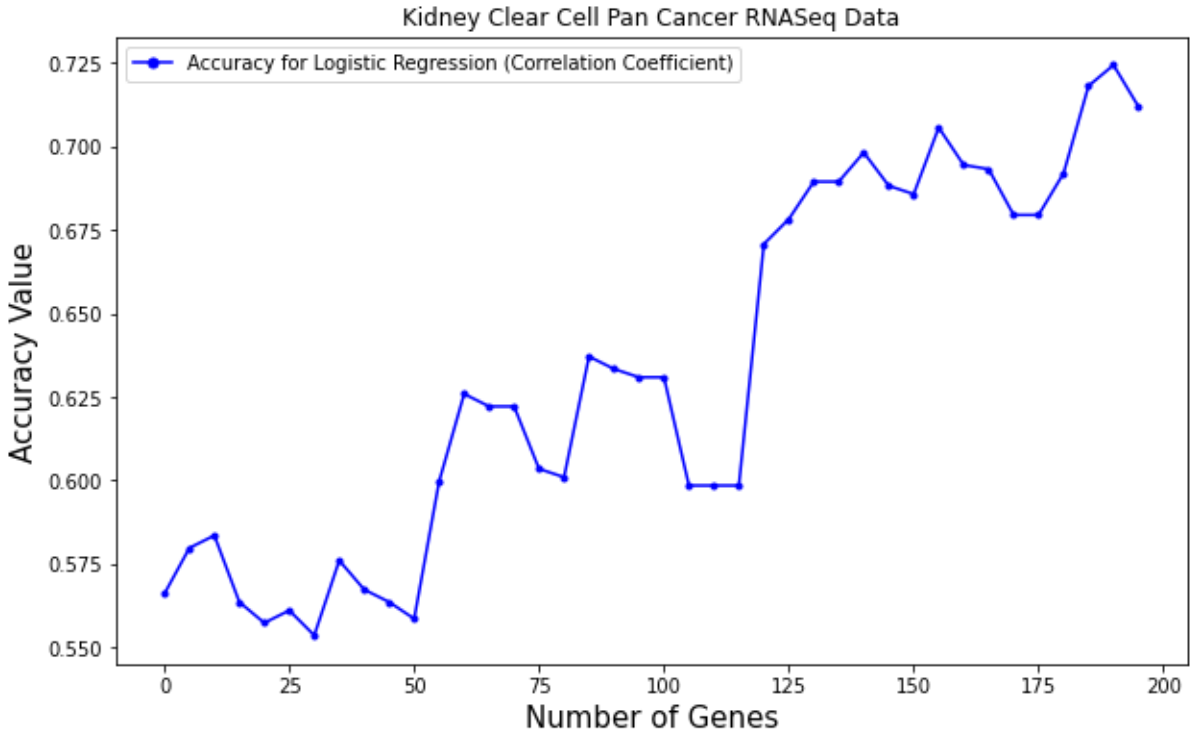


Figure 2: Accuracy against Number of Genes for TCGA PanCancer Atlas Dataset

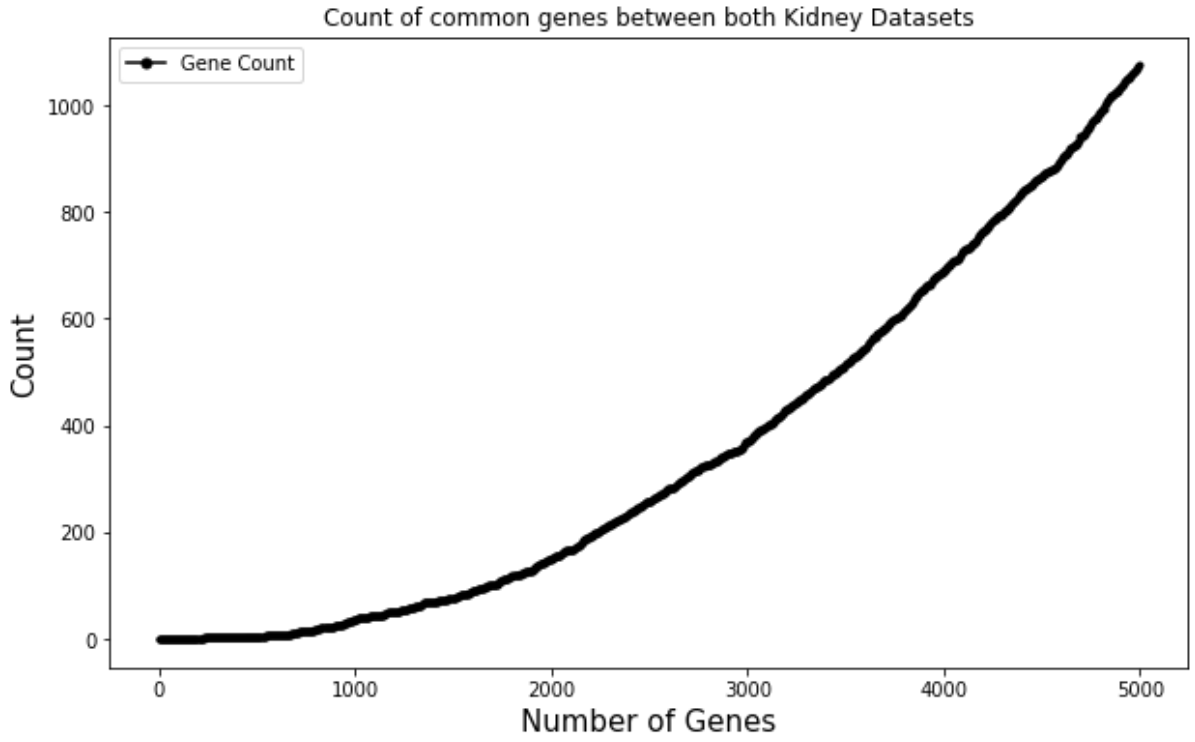


Figure 3: Gene Count against Number of Genes Common between both Datasets

Discussion:

The results in Figure 1 and Figure 2 show an upward trend as expected. As more genes (features) are added sequentially, the accuracy increases up to 75% and 72% for Firehose Legacy and PanCancer Atlas Datasets respectively. The correlation coefficient values are in Table 2.

The number of common genes ranked by correlation coefficient magnitude between the two datasets are shown in Figure 3. It shows an exponential increasing trend. The full list of ranked genes are in the excel files attached in the zip file.

Future Work:

Common genes from other datasets such as Colorectal Cancer, Acute Myeloid Leukemia, Bladder Cancer, Diffuse Large B-Cell Lymphoma, Gallbladder Carcinoma and others can be explored.

Other models such as Random Forest and Support Vector Machines can be explored as well. Also, if both microarray and RNA Sequencing data is available, the genes can be compared as well.

Citations:

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), 530-536s