Gene Expression is directly related to the Gleason score

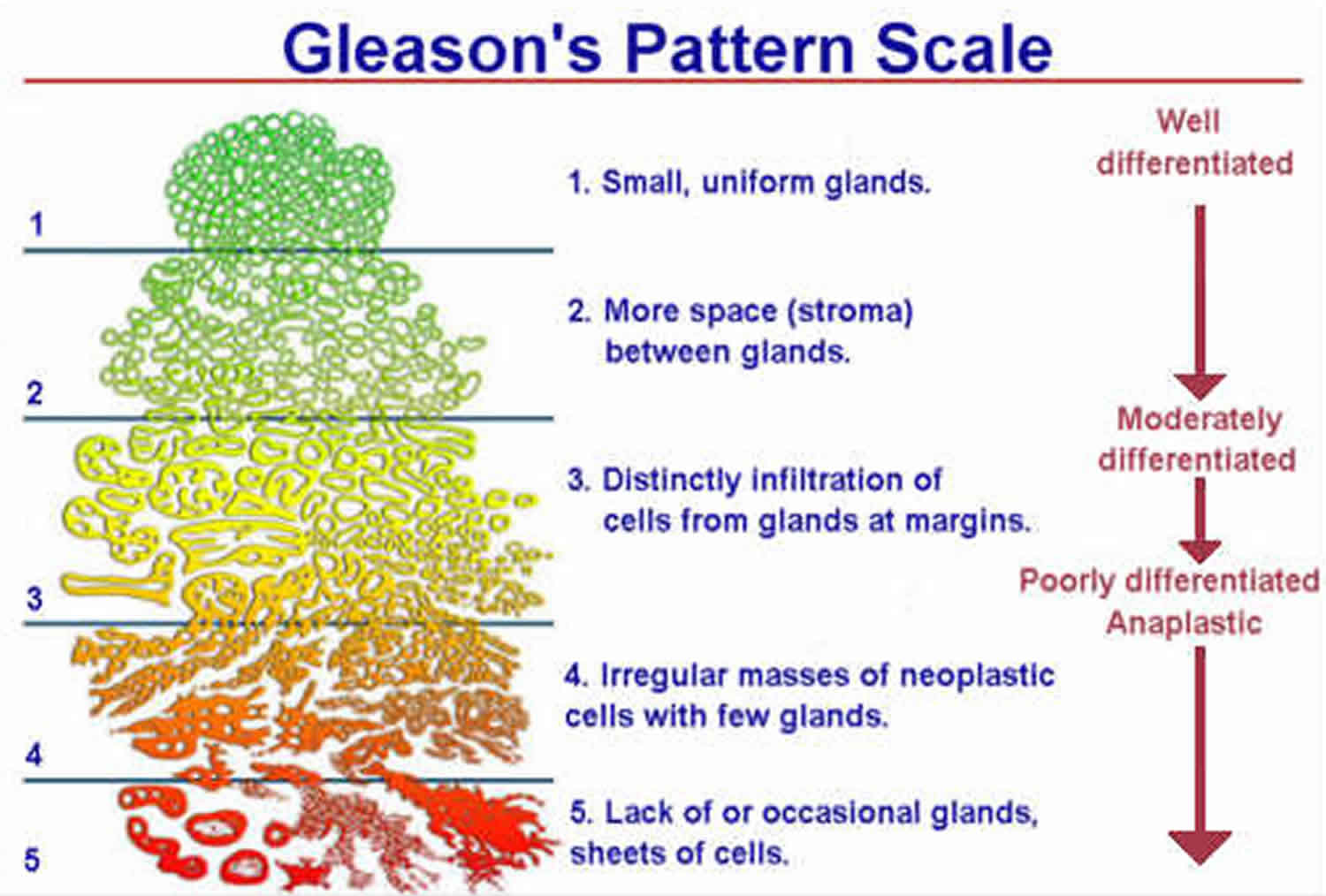
* Some genes will be regulated differently
* We should be able to use a NN to identified the genes involved by inspecting the weights and bias of the network

**Why not DEGs?**

There are so many overlaps between the gene expression patterns.

**Gleason grading system**

The Gleason score is the grading system used to determine the aggressiveness of the prostate cancer. This grading system can be used to choose appropriate treatment options.



Structure of Dataset:

Gene - row data

Sample – column data

Gene expression data stored in assay

Removing Metastatic (prostate cancer) because there is only one ampng 553

Download data from TCGA

Construction of neural network (MLP )

Cleaning data and structuring of data

Extraction of genes based on weights and bias of layer 1 nodes

Extraction of model weights

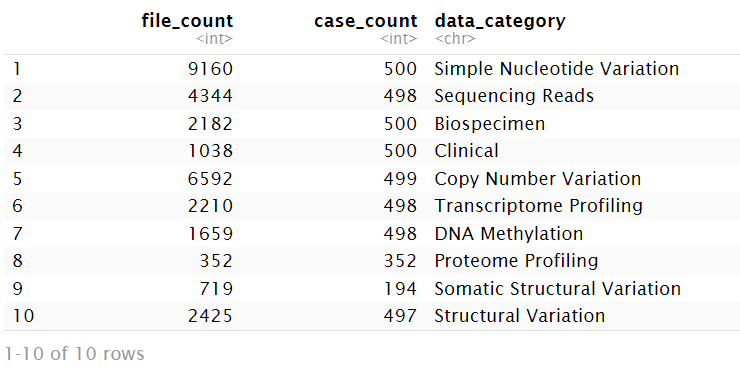
Training of Neural Networks

Gene set enrichment of GOI

Conversion of Gene ID

Fig. 2: Methodology

Importing data from TCGA



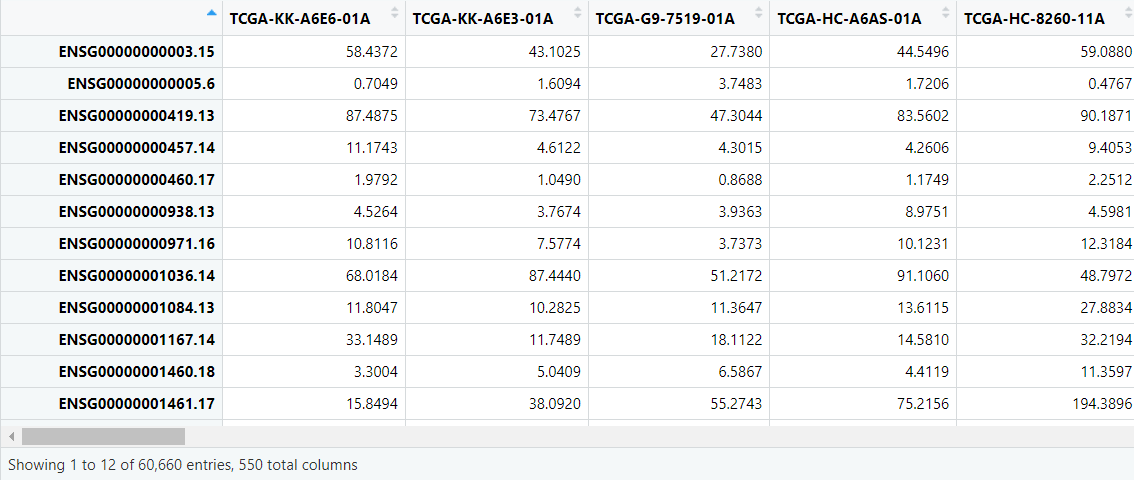
Genes that have been expressed more than 2000 tpm if gene expression is too high most likely to housekeeping gene and they are not useful to differentiate cancer gleason score.

Genes that have account rate less than 100 maximum , too low so that they are not going to contribution to neural network.

5909 gene with 500 patients

Normalization:

Train data



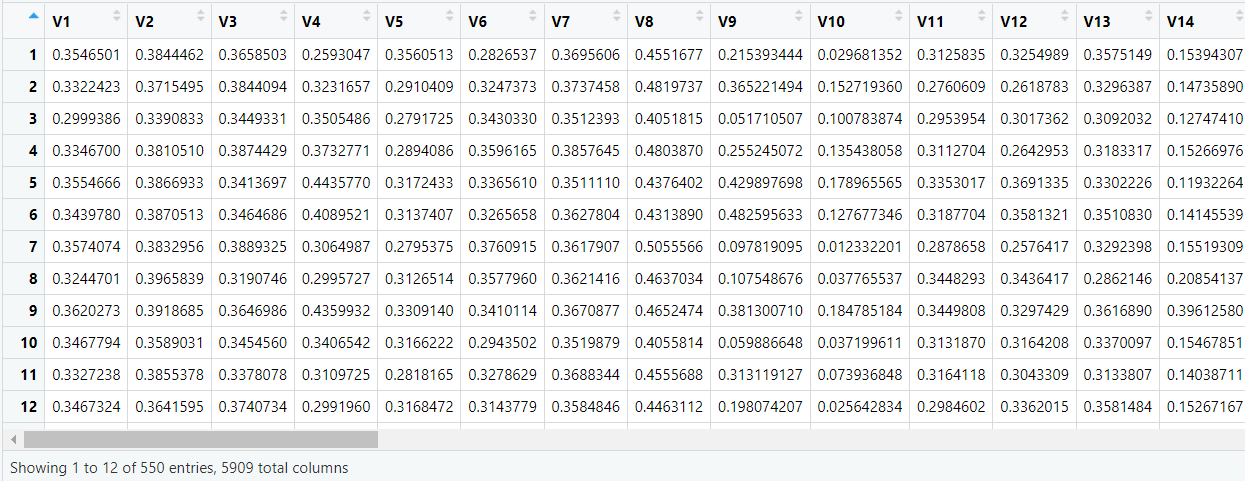
**Min max normalization and Log2 normalization**

Min max normalization so that the lowest count rate in sample is assigned as 0 and highest cal rate is assigned as 1 so everything will go between 0 and 1 .

Log2 normalization so that data is normally distributed otherwise for gene expression they are more skilled to negative side so it is not that great.

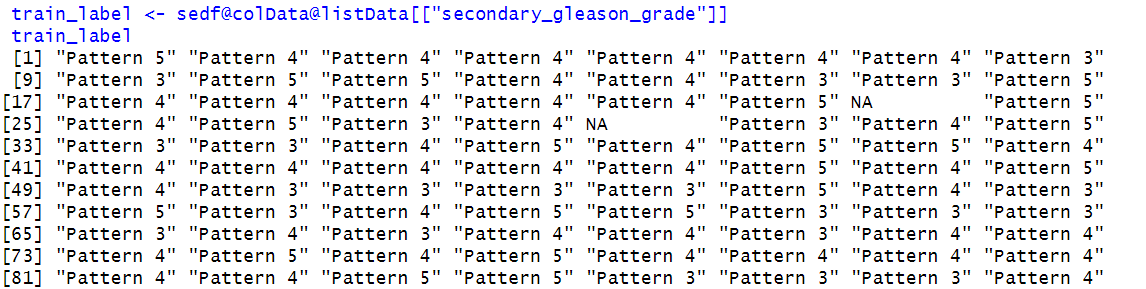
While transpose :

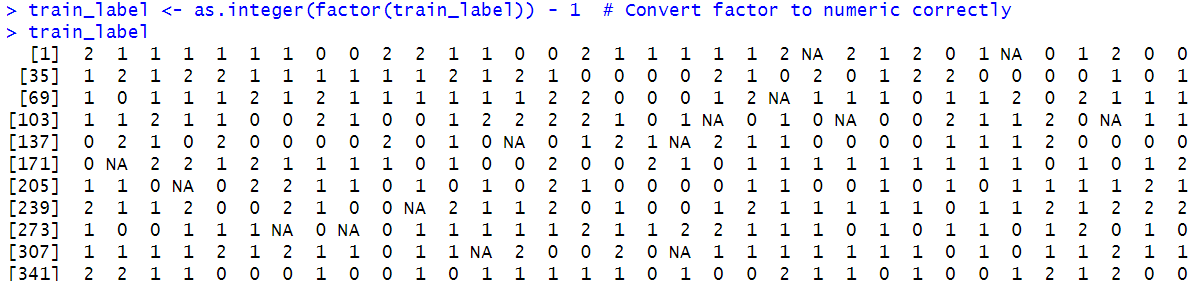
Gene become and sample becomes row

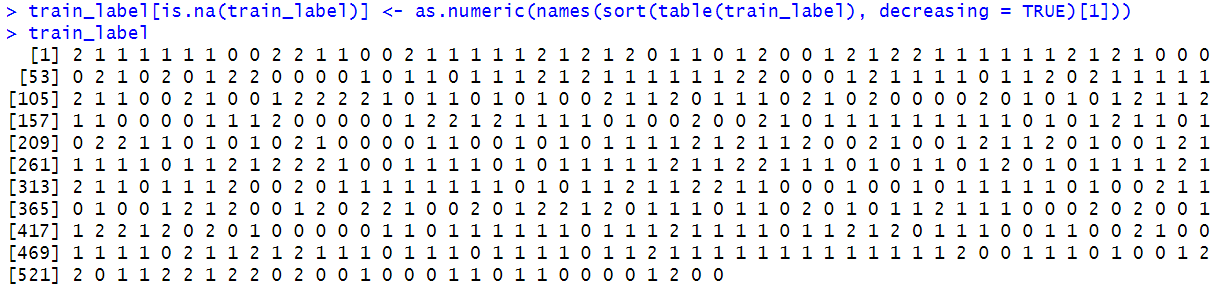


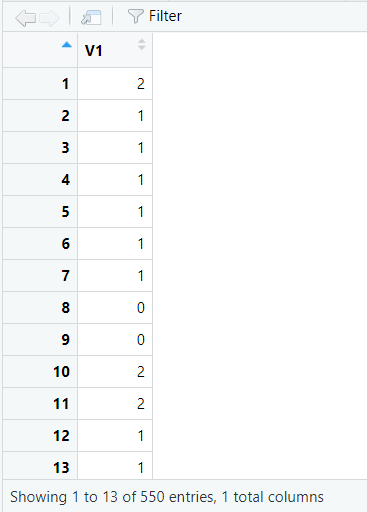
From PCA we can see that the data is categorical data

Data preparation:









The function to\_categorical(train\_label,num\_classes=3) is used for one-hot encoding the categorical labels in the dataset.



Indicating 0,1,2 into V1,V2,V3

Creating heatmap

