1. **What I have done:**
2. Did multiclass classification on the cancer types with features set as sbs signatures, using the softmax to construct the BPNet to perform the training and testing of the model, the result of the classification accuracy pf all the fold’s validation dataset is of

[0.7875739644970414,0.7473372781065089,0.7437869822485207,0.7366863905325444, 0.7461538461538462]

1. made the new cross validation strategy to test on generalization of the model.

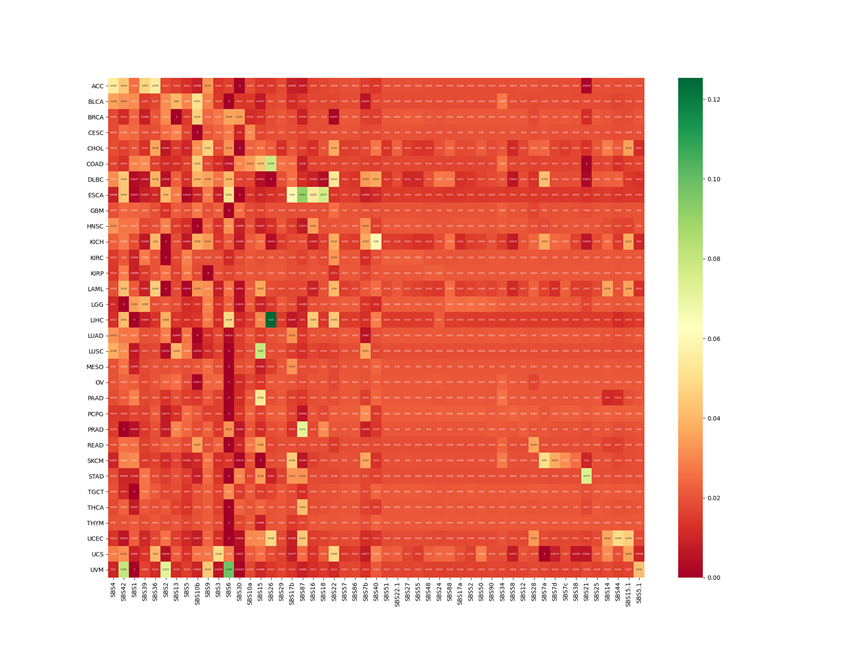
The graph below shows the working flow.

The approach is to separate the whole data into 6 dataset (5 cross validation dataset and 1 validation dataset) in each iterations of 5 iterations, we select the ith dataset in 5 cross validation dataset as testing set and rest as training set using the BPnet and at each iteration, when evaluate the best trained model on the evaluation set

Diagram, schematic

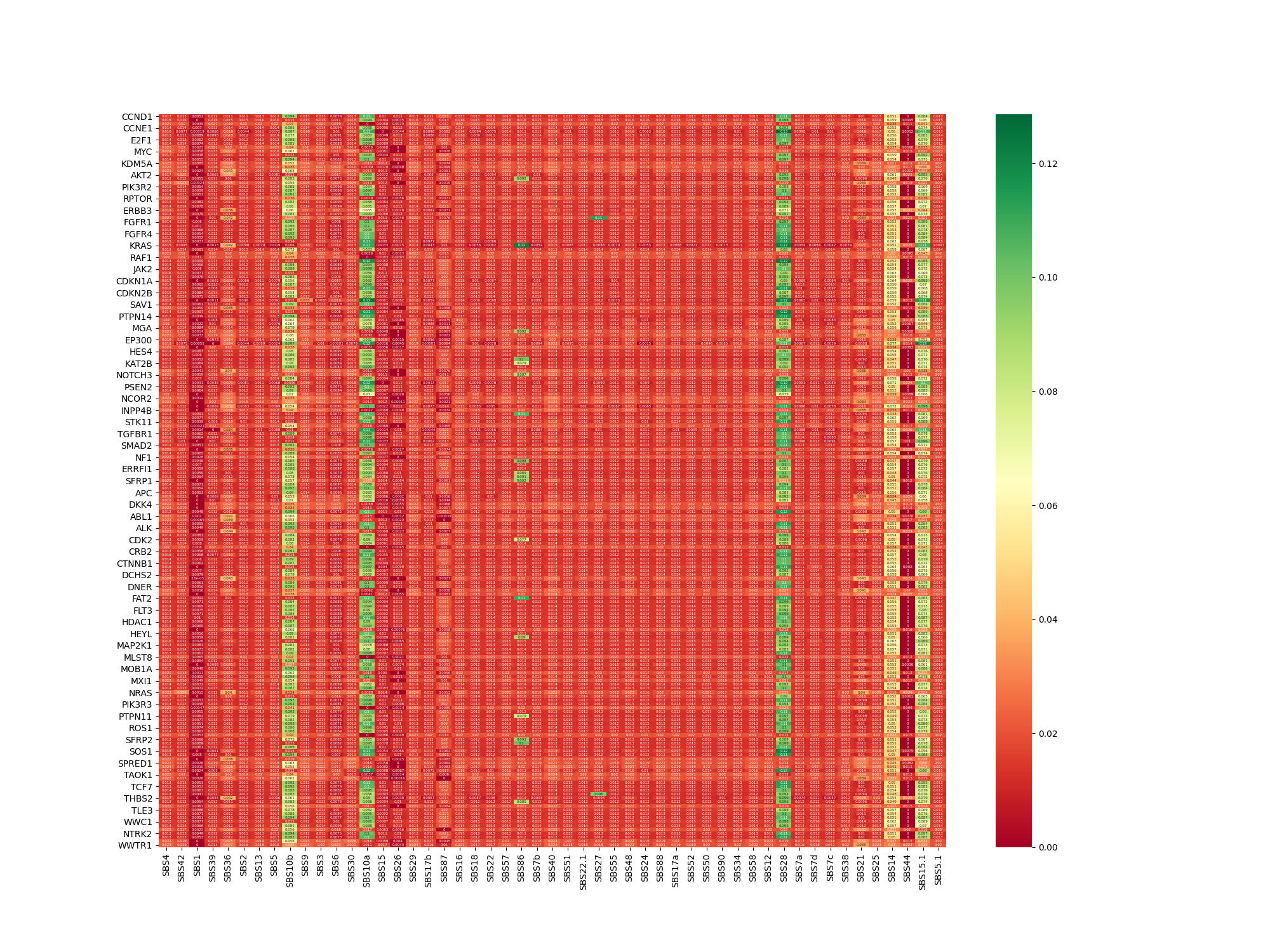
Description automatically generated

This is for the weight of the sbs signature in each cancer types.

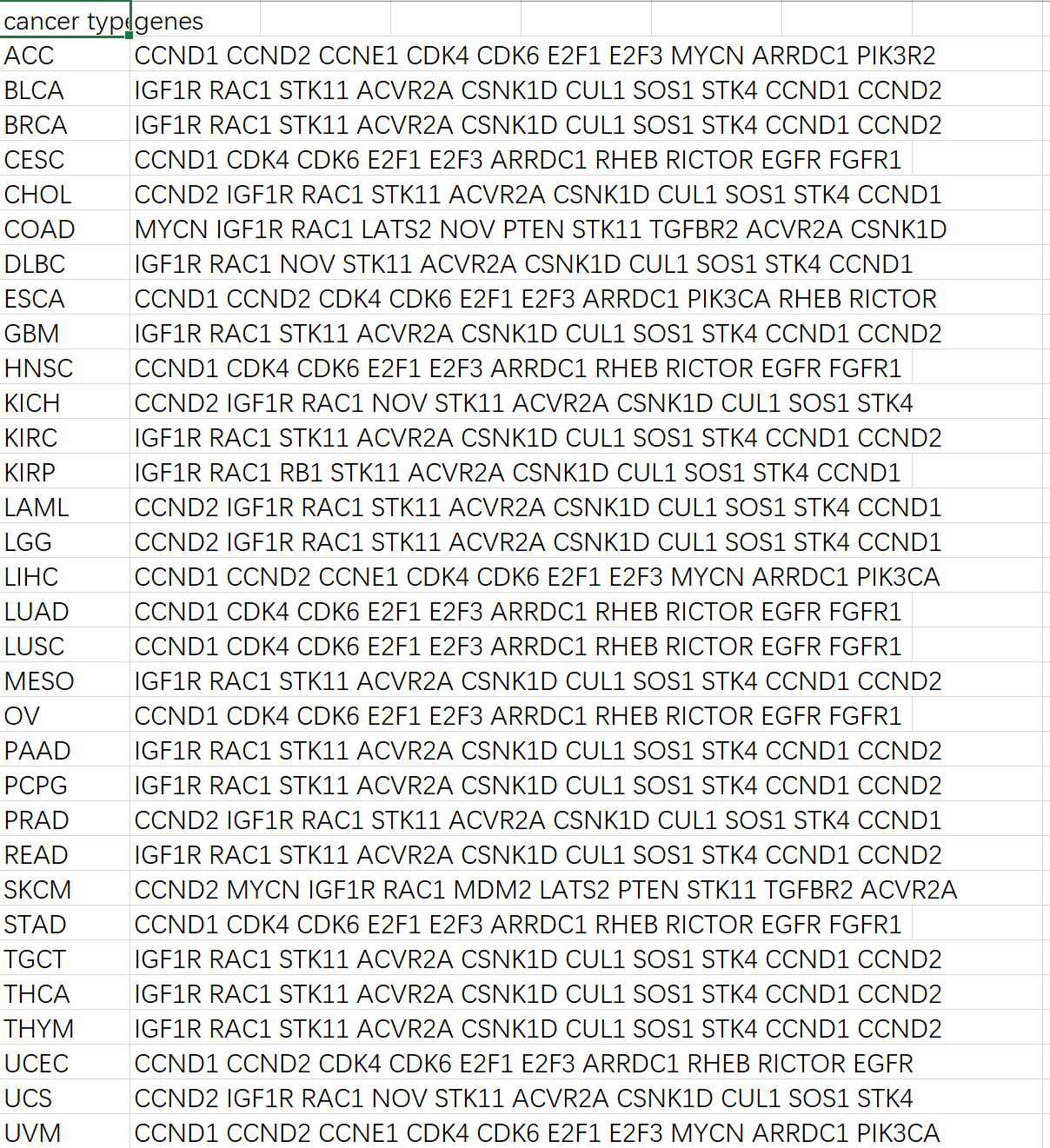


1. Did the classification on the gene mutation status with features set as sbs signatures, using the softmax to construct the BPNet to perform the training and testing of the model, the result of the classification is really good as it has only the accuracy of 97%,haven’t done the evaluation on it yet as I want to make sure with supervisor about the technique of doing cross validation.

This is for the sbs weight in each genes



4.found the heavier weight of sbs in each cancer type and heavier weight of each sbs in each gene and taken the length of the intersection set of those sbs signatures and divided by length of total sbs signatures to find the intersection (cancer type & gene set) that has most of the sbs signatures covered, which means that this gene's mutation has most likely caused the occurrence of heavier sbs signatures used to identify the cancer type. thus, found the top 10 most determinable genes (the activity of those gene might cause the specific cancer) in each cancer types. The graph below shows the top 10 genes in each cancer types.



So the graph of explaining what did I do in finding the relationship between gene and cancer types is presented here:

The most weighted sbs signatures for different cancer types

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACC | Sbs1 | Sbs2 | Sbs3 | Sbs4 | Sbs5 | Sbs6 | Sbs7 | Sbs8 |
| LGG | Sbs2 | Sbs3 | Sbs4 | Sbs5 | Sbs6 | Sbs7 | Sbs8 | Sbs9 |
| UVM | Sbs17 | Sbs67 | Sbs89 | Sbs90 | Sbs2 | Sbs5 | Sbs7 | Sbs8 |

The most weighted sbs signatures for different genes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene1 | Sbs1 | Sbs2 | Sbs3 | Sbs4 | Sbs5 | Sbs6 |  |  |
| Gene2 | Sbs1 | Sbs2 |  |  |  |  |  |  |

So, the common set for ACC and gene1 would be sbs1,2,3,4,5,6, assuming there’s 52 sbs signatures in total, we calculate the gene in cancer as

Relation of ACC and gen1 : num(sbs1,2,3,4,5,6)/len(all sbs) = 6/52

Relation of ACC and gene2 : num(sbs1,2)/len(all sbs) = 2/52

Then,the relationship between gene1 and ACC is much denser than that of gene2 and ACC.we list top 10 of the genes in each of the cancers an relate them.

4.Wrote the data and data preprocessing part of dissertation, need to find the reference to make sure the reason why it’s using sbs as features is explained

**2.The goal:**

1. we look into whether the sbs signatures could be used to determine the cancer types
2. we then find the most powerful sbs signatures for determine such cancer (eg. Lung cancer), then we set those sbs signatures as input and then set the gene mutation status as labels and classify the gene mutation status.
3. To be able to know the mutation status of the genes without sequencing the tumor, we could build recommendation system to predict which of the gene are mutated in different cancer types.

**3.The problems:**

1. Do you want to predict the gene mutation status like predicting it as 0 or 1 by just setting cancer types as x ? or do you just want to know the weight of each gene **(the most important)**in each cancer types.
2. I want to ensure again that what I am only expected to do is to analyze the classification of cancer types using sbs signatures and the analysis of the classification and prediction of gene mutation status using sbs signatures and find the connection between cancer type and gene mutation status?

**4.The plan**

(1). discuss the problem of the prediction of the gene mutation status with supervisor!

(2) adding the evaluation model to gene classification

(3) use cancer type to predict genes mutation status

(3). working on method part of the dissertation