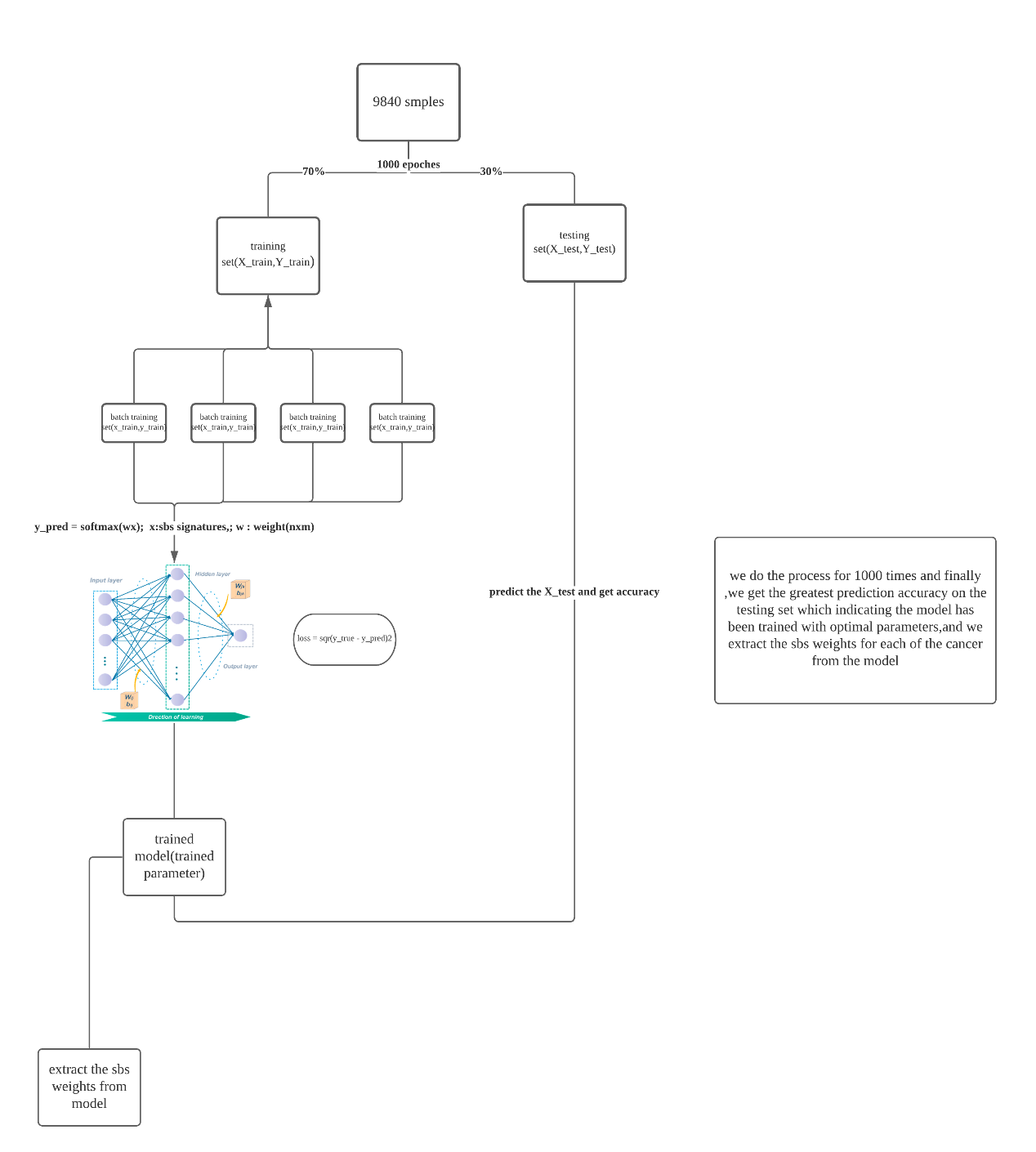
1. **What I have done:**
2. Did multiclass classification on the cancer types with features set as sbs signatures, using the softmax to construct the BPNet to perform the training and testing of the model, the result of the classification is really good as it has only the accuracy of 93.223%

The graph below shows the working flow.



The classification report.

precision recall f1-score support

0 0.73 0.97 0.83 36

1 1.00 0.98 0.99 124

2 0.94 0.98 0.96 300

3 0.95 0.96 0.95 76

4 0.00 0.00 0.00 10

5 0.99 0.98 0.99 115

6 0.00 0.00 0.00 14

7 0.85 0.98 0.91 47

8 0.91 0.91 0.91 130

9 0.93 0.99 0.96 144

10 0.00 0.00 0.00 21

11 0.92 1.00 0.96 97

12 0.97 0.88 0.93 77

13 0.00 0.00 0.00 30

14 1.00 0.96 0.98 138

15 0.99 1.00 1.00 109

16 0.99 0.95 0.97 193

17 1.00 0.97 0.98 157

18 0.77 0.92 0.84 26

19 0.93 0.98 0.96 128

20 0.86 0.89 0.88 55

21 0.55 1.00 0.71 54

22 0.99 0.95 0.97 143

23 0.91 1.00 0.95 40

24 0.98 0.98 0.98 120

25 0.98 0.95 0.96 130

26 0.76 0.98 0.86 52

27 0.92 0.99 0.95 149

28 0.49 0.51 0.50 35

29 0.99 0.99 0.99 157

30 0.00 0.00 0.00 22

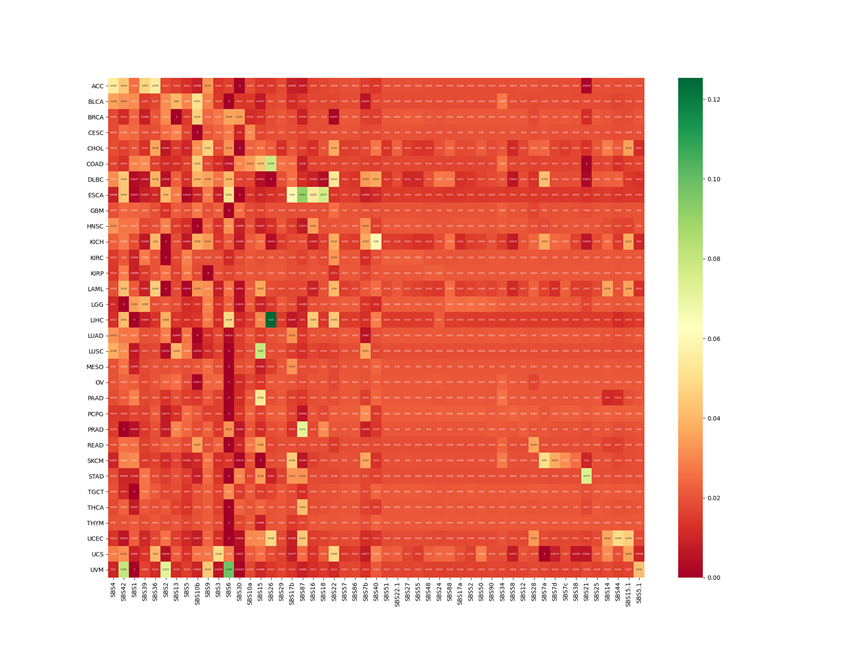
31 1.00 0.68 0.81 22

accuracy 0.93 2951

macro avg 0.76 0.79 0.77 2951

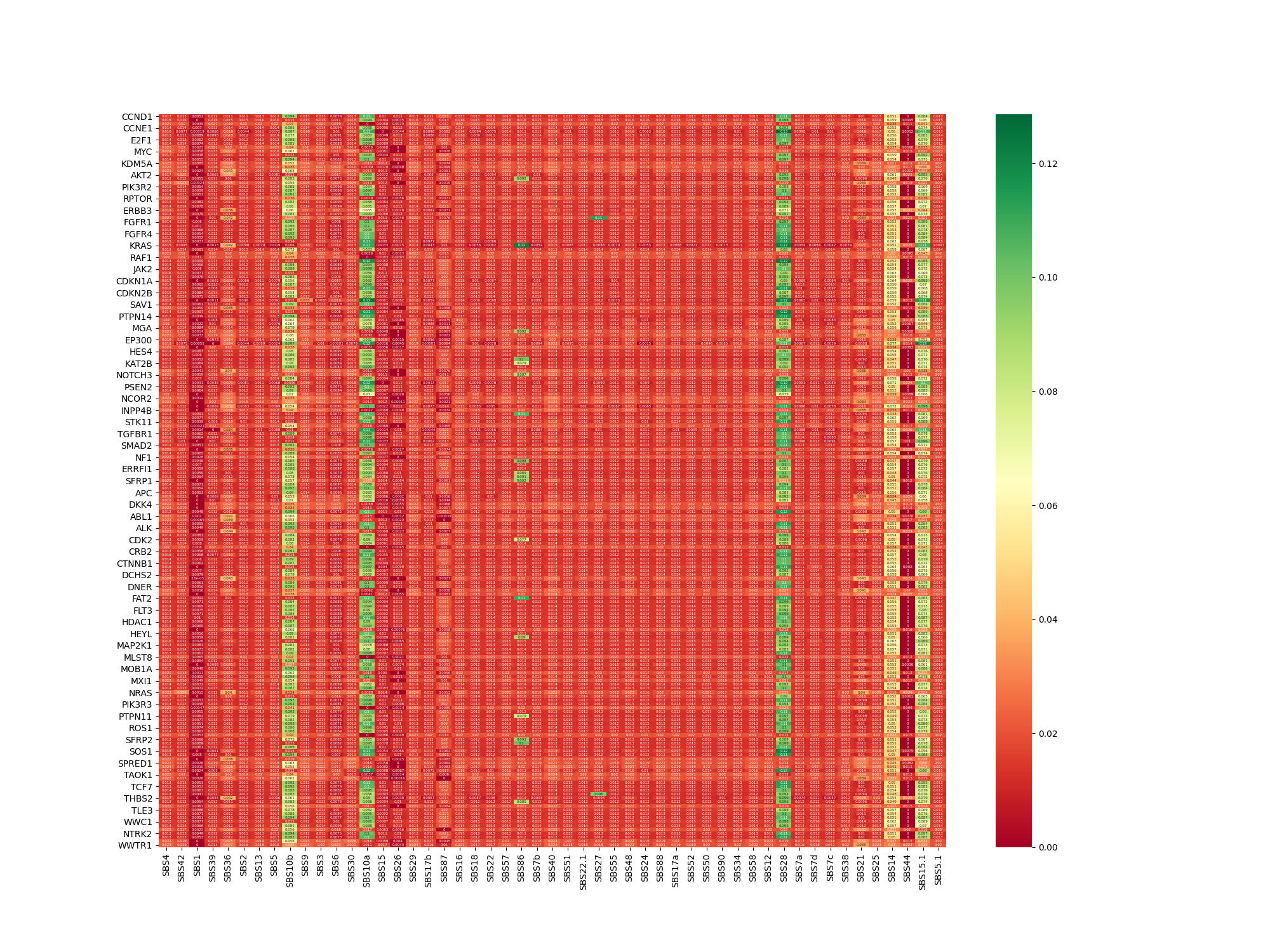
weighted avg 0.91 0.93 0.92 2951

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

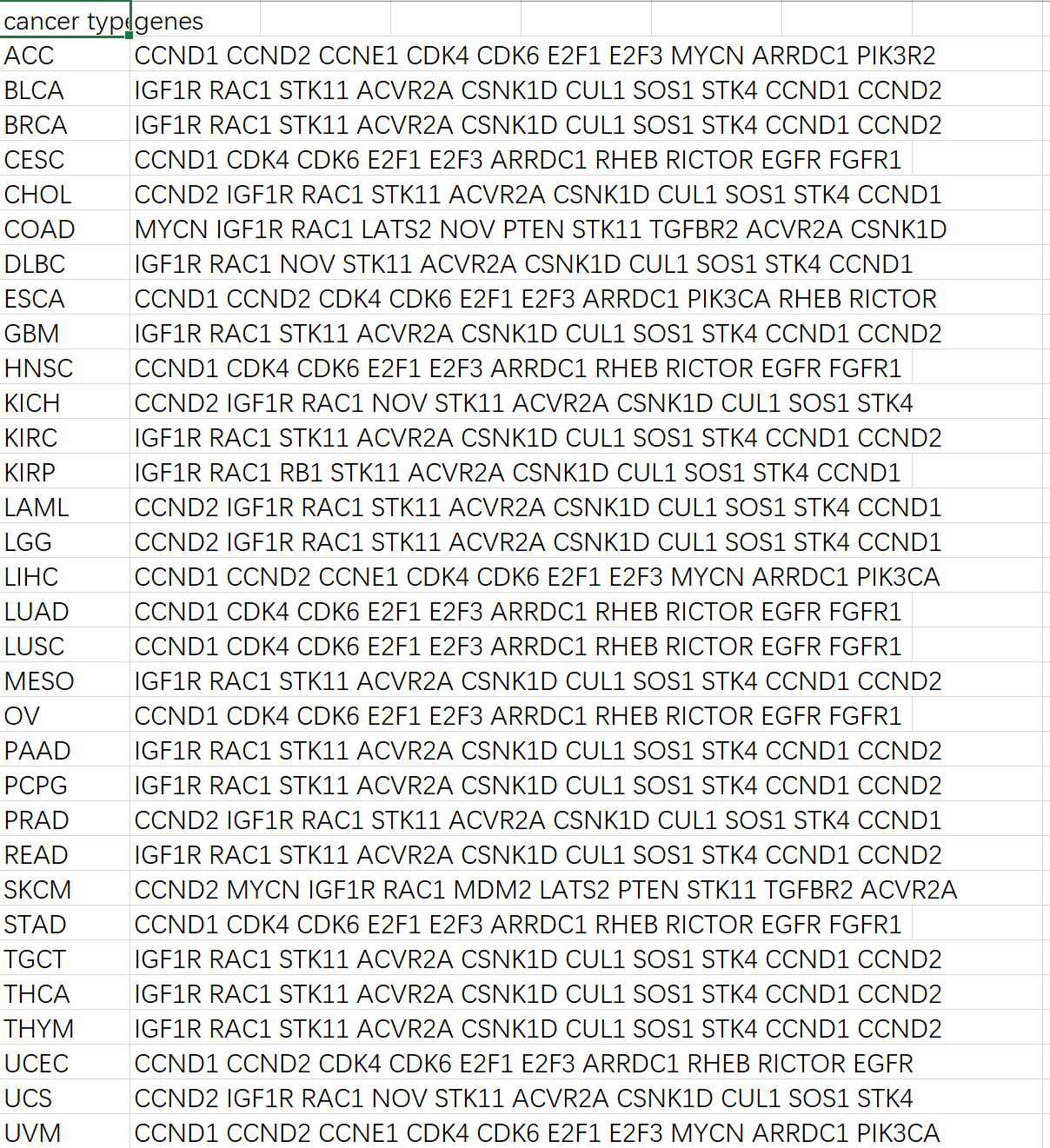


1. Did the classification on the gene mutation status with features set as sbs signatures, using the softmax to construct the BPNet to perform the training and testing of the model, the result of the classification is really good as it has only the accuracy of 97%

This is for the sbs weight in each genes



3.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



4.Wrote the data and data preprocessing part of dissertation but it’s still in modification.

**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 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.

If we want to be able to predict the gene mutation status,would those be good idea?

(do yo think taking the heavier weight sbs in cancer classification for each cancer types and using those sbs to classify and predict gene mutation status is a good idea or we use the intersection of the sbs to predict the gene mutation status?)

1. As for the gene relations ,do you know where I can find related articles.

3, for the evaluation part of the project,even through the prediction is accurate for all of the sample data, but what if its not accurate in future prediction ? we don’t have any doctor to verify the prediction of gene mutation status in each cancer types

**4.The plan**

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

(2). summarize the plan and decide whether to continuing constructing recommender system on gene mutation status for related genes.

(3). working on method part of the dissertation