1. **What I have done:**

(1).

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, and extracted weight of each sbs signatures in each cancer types and gene types. **the result of the classification accuracy of all the fold's validation dataset for cancer is of (5-fold cross validation):**

The 5-fold cross validation has 5 testing result, they are:

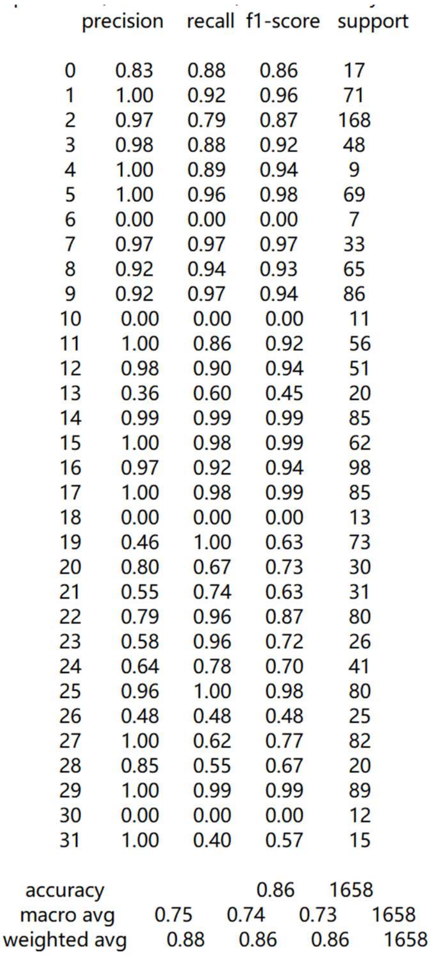
[0.8809523809523809, **0.8897243107769424**, 0.8784461152882206, 0.8515037593984962, 0.8602756892230576]

The validation accuracies for 5-fold cross validation are:

[**0.8600723763570567**, 0.850422195416164, 0.8528347406513872, 0.8462002412545235, 0.853437876960193]

The classification report for the classification status of 5 validation sets are shown below:

The first time for the validation set: **86%**, The cancers have 0 accuracy are 6,10,18,30.

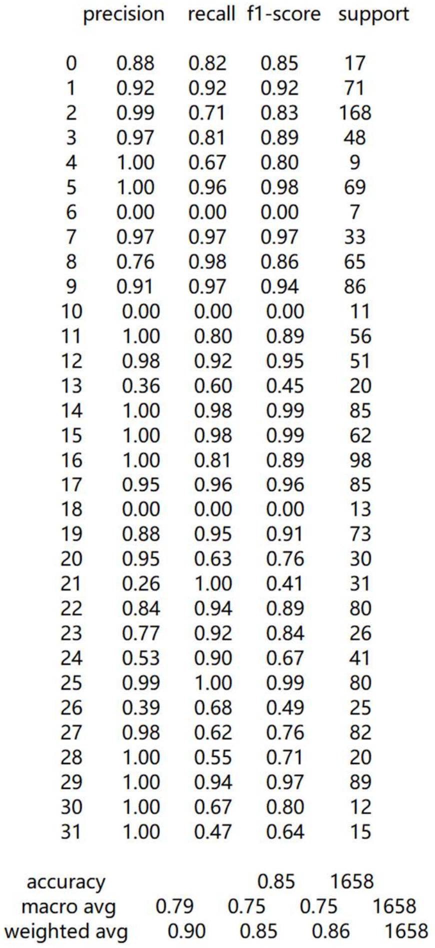


```
          precision    recall  f1-score   support

       0       0.83      0.88      0.86        17
       1       1.00      0.92      0.96        71
       2       0.97      0.79      0.87       168
       3       0.98      0.88      0.92        48
       4       1.00      0.89      0.94         9
       5       1.00      0.96      0.98        69
       6       0.00      0.00      0.00         7
       7       0.97      0.97      0.97        33
       8       0.92      0.94      0.93        65
       9       0.92      0.97      0.94        86
      10       0.00      0.00      0.00        11
      11       1.00      0.86      0.92        56
      12       0.98      0.90      0.94        51
      13       0.36      0.60      0.45        20
      14       0.99      0.99      0.99        85
      15       1.00      0.98      0.99        62
      16       0.97      0.92      0.94        98
      17       1.00      0.98      0.99        85
      18       0.00      0.00      0.00        13
      19       0.46      1.00      0.63        73
      20       0.80      0.67      0.73        30
      21       0.55      0.74      0.63        31
      22       0.79      0.96      0.87        80
      23       0.58      0.96      0.72        26
      24       0.64      0.78      0.70        41
      25       0.96      1.00      0.98        80
      26       0.48      0.48      0.48        25
      27       1.00      0.62      0.77        82
      28       0.85      0.55      0.67        20
      29       1.00      0.99      0.99        89
      30       0.00      0.00      0.00        12
      31       1.00      0.40      0.57        15

    accuracy                           0.86      1658
   macro avg       0.75      0.74      0.73      1658
weighted avg       0.88      0.86      0.86      1658
```

The second time: **85%, **The cancers have 0 accuracy are 6,10,18

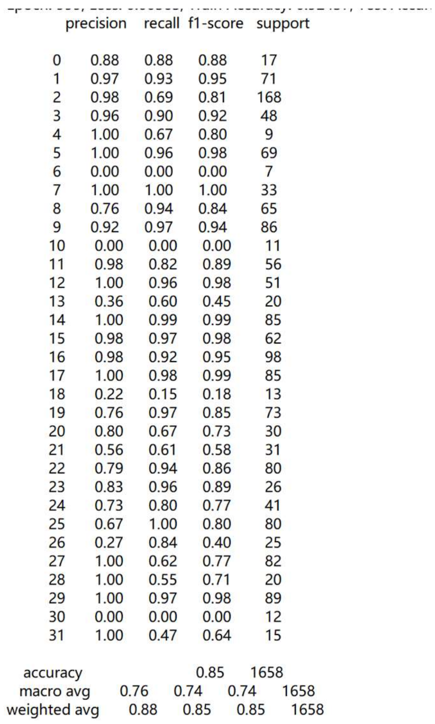


```
             precision    recall  f1-score   support

           0       0.88      0.82      0.85        17
           1       0.92      0.92      0.92        71
           2       0.99      0.71      0.83       168
           3       0.97      0.81      0.89        48
           4       1.00      0.67      0.80         9
           5       1.00      0.96      0.98        69
           6       0.00      0.00      0.00         7
           7       0.97      0.97      0.97        33
           8       0.76      0.98      0.86        65
           9       0.91      0.97      0.94        86
          10       0.00      0.00      0.00        11
          11       1.00      0.80      0.89        56
          12       0.98      0.92      0.95        51
          13       0.36      0.60      0.45        20
          14       1.00      0.98      0.99        85
          15       1.00      0.98      0.99        62
          16       1.00      0.81      0.89        98
          17       0.95      0.96      0.96        85
          18       0.00      0.00      0.00        13
          19       0.88      0.95      0.91        73
          20       0.95      0.63      0.76        30
          21       0.26      1.00      0.41        31
          22       0.84      0.94      0.89        80
          23       0.77      0.92      0.84        26
          24       0.53      0.90      0.67        41
          25       0.99      1.00      0.99        80
          26       0.39      0.68      0.49        25
          27       0.98      0.62      0.76        82
          28       1.00      0.55      0.71        20
          29       1.00      0.94      0.97        89
          30       1.00      0.67      0.80        12
          31       1.00      0.47      0.64        15

    accuracy                           0.85      1658
   macro avg       0.79      0.75      0.75      1658
weighted avg       0.90      0.85      0.86      1658
```

The third time: **85%**, The cancers have 0 accuracy are 6,10,30.



```
             precision    recall  f1-score   support

           0       0.88      0.88      0.88        17
           1       0.97      0.93      0.95        71
           2       0.98      0.69      0.81       168
           3       0.96      0.90      0.92        48
           4       1.00      0.67      0.80         9
           5       1.00      0.96      0.98        69
           6       0.00      0.00      0.00         7
           7       1.00      1.00      1.00        33
           8       0.76      0.94      0.84        65
           9       0.92      0.97      0.94        86
          10       0.00      0.00      0.00        11
          11       0.98      0.82      0.89        56
          12       1.00      0.96      0.98        51
          13       0.36      0.60      0.45        20
          14       1.00      0.99      0.99        85
          15       0.98      0.97      0.98        62
          16       0.98      0.92      0.95        98
          17       1.00      0.98      0.99        85
          18       0.22      0.15      0.18        13
          19       0.76      0.97      0.85        73
          20       0.80      0.67      0.73        30
          21       0.56      0.61      0.58        31
          22       0.79      0.94      0.86        80
          23       0.83      0.96      0.89        26
          24       0.73      0.80      0.77        41
          25       0.67      1.00      0.80        80
          26       0.27      0.84      0.40        25
          27       1.00      0.62      0.77        82
          28       1.00      0.55      0.71        20
          29       1.00      0.97      0.98        89
          30       0.00      0.00      0.00        12
          31       1.00      0.47      0.64        15

    accuracy                           0.85      1658
   macro avg       0.76      0.74      0.74      1658
weighted avg       0.88      0.85      0.85      1658
```

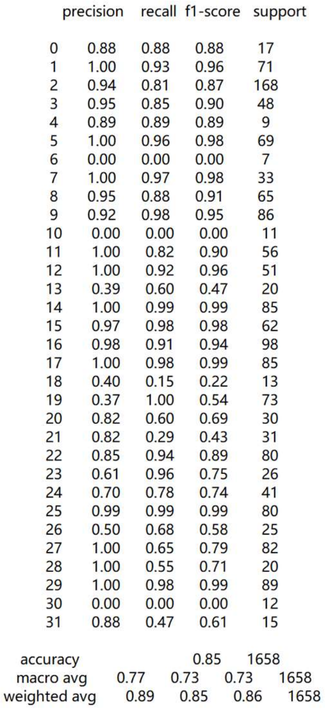
The fourth time: **85%**, The cancers have 0 accuracy are 6,10,18,30.

```
              precision    recall  f1-score   support

           0       0.88      0.88      0.88        17
           1       1.00      0.92      0.96        71
           2       1.00      0.68      0.81       168
           3       0.98      0.83      0.90        48
           4       1.00      0.89      0.94         9
           5       1.00      0.97      0.99        69
           6       0.00      0.00      0.00         7
           7       0.97      0.97      0.97        33
           8       0.76      0.97      0.85        65
           9       0.94      0.95      0.95        86
          10       0.00      0.00      0.00        11
          11       1.00      0.75      0.86        56
          12       0.98      0.92      0.95        51
          13       0.36      0.60      0.45        20
          14       1.00      0.99      0.99        85
          15       0.95      1.00      0.98        62
          16       0.99      0.90      0.94        98
          17       0.99      0.98      0.98        85
          18       0.00      0.00      0.00        13
          19       0.62      1.00      0.77        73
          20       0.44      0.83      0.57        30
          21       0.57      0.55      0.56        31
          22       1.00      0.90      0.95        80
          23       0.69      0.96      0.81        26
          24       0.70      0.76      0.73        41
          25       0.98      0.99      0.98        80
          26       0.19      0.76      0.31        25
          27       1.00      0.62      0.77        82
          28       1.00      0.55      0.71        20
          29       1.00      0.99      0.99        89
          30       0.00      0.00      0.00        12
          31       1.00      0.47      0.64        15

    accuracy                           0.85      1658
   macro avg       0.75      0.74      0.72      1658
weighted avg       0.89      0.85      0.85      1658
```

The fifth time: **85%**, The cancers have 0 accuracy are 6,10, 30.



```
              precision    recall  f1-score   support

           0       0.88      0.88      0.88        17
           1       1.00      0.93      0.96        71
           2       0.94      0.81      0.87       168
           3       0.95      0.85      0.90        48
           4       0.89      0.89      0.89         9
           5       1.00      0.96      0.98        69
           6       0.00      0.00      0.00         7
           7       1.00      0.97      0.98        33
           8       0.95      0.88      0.91        65
           9       0.92      0.98      0.95        86
          10       0.00      0.00      0.00        11
          11       1.00      0.82      0.90        56
          12       1.00      0.92      0.96        51
          13       0.39      0.60      0.47        20
          14       1.00      0.99      0.99        85
          15       0.97      0.98      0.98        62
          16       0.98      0.91      0.94        98
          17       1.00      0.98      0.99        85
          18       0.40      0.15      0.22        13
          19       0.37      1.00      0.54        73
          20       0.82      0.60      0.69        30
          21       0.82      0.29      0.43        31
          22       0.85      0.94      0.89        80
          23       0.61      0.96      0.75        26
          24       0.70      0.78      0.74        41
          25       0.99      0.99      0.99        80
          26       0.50      0.68      0.58        25
          27       1.00      0.65      0.79        82
          28       1.00      0.55      0.71        20
          29       1.00      0.98      0.99        89
          30       0.00      0.00      0.00        12
          31       0.88      0.47      0.61        15

    accuracy                           0.85      1658
   macro avg       0.77      0.73      0.73      1658
weighted avg       0.89      0.85      0.86      1658
```

As we can see from the graph above, the overall accuracy is good as the average of them are 85 % and there are still some classes that have the 0 accuracy, there are 2 possible reasons:

1. **The 6th cancer ('DLBC') has number of patients of 7 in each set ,10th ('HNSC') cancer has number of patients of 11 in each set, 18th cancer('LUSC') has number of patients of 13 in each set and 30th ('UCS') has number of patients of 12 in each set.as we can observe from the distribution, those are the cancer types that are minors in the whole data set and therefore, the classification accuracy of theirs are low. The solution would be reducing the batch size to try to help with training or getting more data set, which is time consuming.**

2. **Those cancers are similar to other cancers in the whole dataset and therefore their classification accuracies are interfered by other classes of cancers. The solution would be trying to classify on lesser cancer types and keep those type of cancers to see the accuracy performance of them afterwards.**

Except for that, the performance of multiclass classification is pretty good, and the model is generally tested on each 6-fold and get the close results, also. The validation accuracies are also stable and close to the best trained model's testing accuracy.

The precision: In an imbalanced classification problem with more than two classes, the precision of each class is calculated as the sum of true positives across that classes divided by the sum of true positives and false positives across that classes. Number of identified items are relevant.

The Recall = TP / (TP+FN) (number of relevant items are identified)

The F1-score: The F-score or F-measure is a measure of a test's accuracy. indicating how perfect is the recall and precision.

Support: number of samples in that class

Accuracy: how many items identified are the relevant element.

Marco Average: macro averaging reduces the problem to multiple one-vs-all comparisons. The truth and estimate columns are recoded such that the only two levels are A and other, and then precision is calculated based on those recoded columns. The results of Marco of each class are then weighted together. The **Pr1** is the precision of class 1.its same for recall and f1-score.

$$Pr_{macro} = \frac{Pr_1 + Pr_2 + \ldots + Pr_k}{k} = Pr_1\frac{1}{k} + Pr_2\frac{1}{k} + \ldots + Pr_k\frac{1}{k}$$

Weighted Average: weighted avg is the total number TP (true positive of all classes)/total number of objects in all classes. It's the same for recall and f1-score.

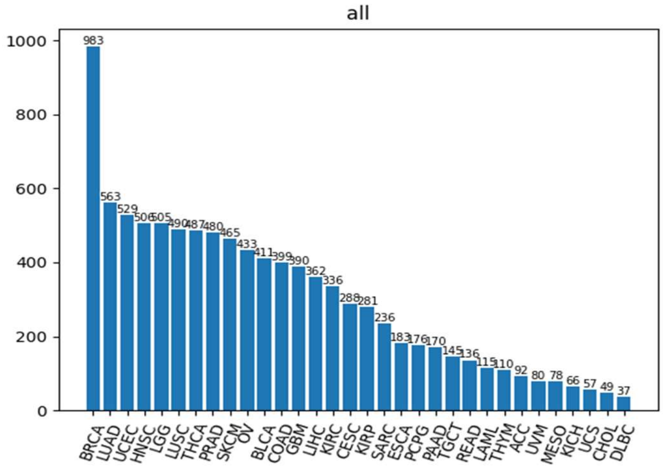
the result of the classification accuracy of all the fold's validation dataset for cancer is of (5-fold cross validation):

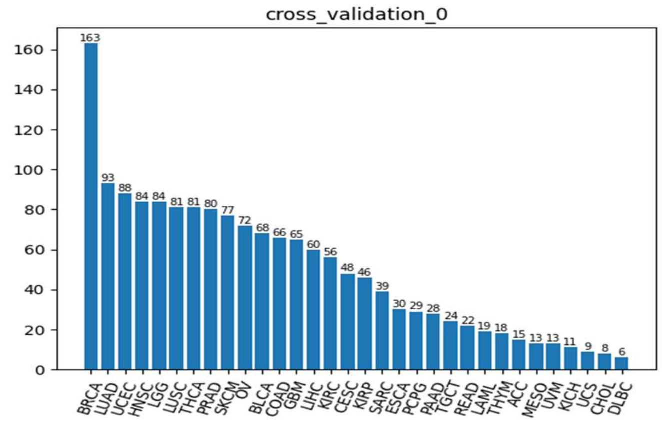
The 5-fold cross validation has 5 testing result, they are: [**0.993734335839599**, 0.9931077694235589,0.9906015037593985, 0.9862155388471178, 0.9874686716791979]

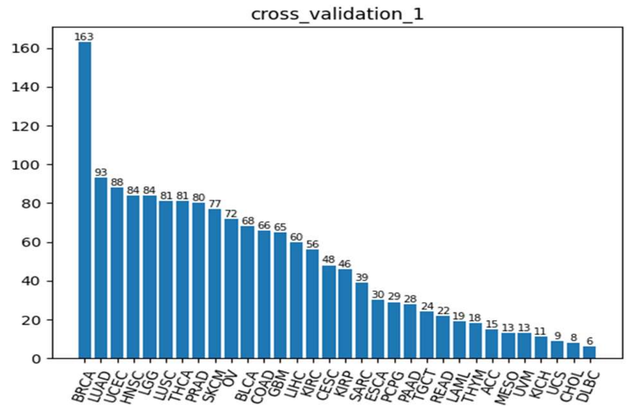
The validation accuracies for 5-fold cross validation are: [0.9927623642943305, **0.9939686369119421**,0.9909529553679132,0.9891435464414958, 0.9909529553679132]

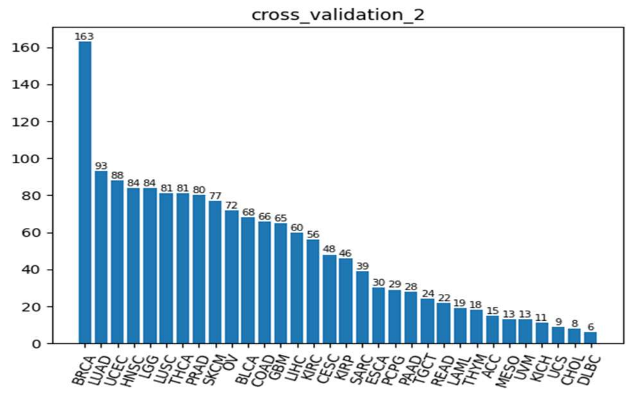
(2).

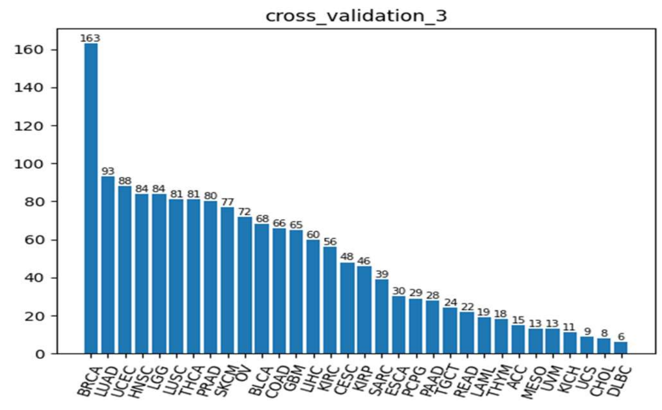
Finished the evaluation stage of making the new cross validation strategy to test on generalization of the model and using the stratified sampling method to ensure the cancer distribution in each cancer types in each folds are the same, here is the cancer distribution in each of the folds.**(Please notice that if you find out that the number of sample in a class is different than the support shown in the above validation classification report is because I have changed the data to test different performance.)**

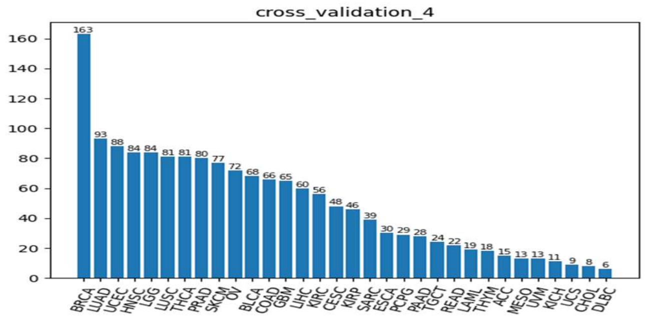


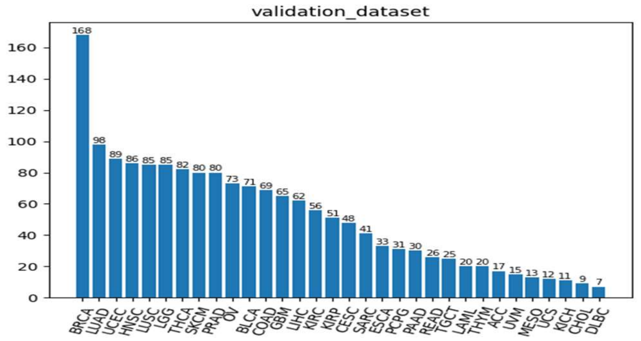

cross_validation_0
0
20
40
60
80
100
120
140
160
163
93
88
84
84
81
81
80
77
72
68
66
65
60
56
48
46
39
30
29
28
24
22
19
18
15
13
13
11
9
8
6
BRCA
LUAD
UCEC
HNSC
LGG
LUSC
THCA
PRAD
SKCM
OV
BLCA
COAD
GBM
LIHC
KIRC
CESC
KIRP
SARC
ESCA
PCPG
PAAD
TGCT
READ
LAML
THYM
ACC
MESO
UVM
KICH
UCS
CHOL
DLBC


cross_validation_1
0
20
40
60
80
100
120
140
160
163
93
88
84
84
81
81
80
77
72
68
66
65
60
56
48
46
39
30
29
28
24
22
19
18
15
13
13
11
9
8
6
BRCA
LUAD
UCEC
HNSC
LGG
LUSC
THCA
PRAD
SKCM
OV
BLCA
COAD
GBM
LIHC
KIRC
CESC
KIRP
SARC
ESCA
PCPG
PAAD
TGCT
READ
LAML
THYM
ACC
MESO
UVM
KICH
UCS
CHOL
DLBC


cross_validation_2
0
20
40
60
80
100
120
140
160
163
93
88
84
84
81
81
80
77
72
68
66
65
60
56
48
46
39
30
29
28
24
22
19
18
15
13
13
11
9
8
6
BRCA
LUAD
UCEC
HNSC
LGG
LUSC
THCA
PRAD
SKCM
OV
BLCA
COAD
GBM
LIHC
KIRC
CESC
KIRP
SARC
ESCA
PCPG
PAAD
TGCT
READ
LAML
THYM
ACC
MESO
UVM
KICH
UCS
CHOL
DLBC


cross_validation_3
0
20
40
60
80
100
120
140
160
163
93
88
84
84
81
81
80
77
72
68
66
65
60
56
48
46
39
30
29
28
24
22
19
18
15
13
13
11
9
8
6
BRCA
LUAD
UCEC
HNSC
LGG
LUSC
THCA
PRAD
SKCM
OV
BLCA
COAD
GBM
LIHC
KIRC
CESC
KIRP
SARC
ESCA
PCPG
PAAD
TGCT
READ
LAML
THYM
ACC
MESO
UVM
KICH
UCS
CHOL
DLBC


cross_validation_4
0
20
40
60
80
100
120
140
160
163
93
88
84
84
81
81
80
77
72
68
66
65
60
56
48
46
39
30
29
28
24
22
19
18
15
13
13
11
9
8
6
BRCA
LUAD
UCEC
HNSC
LGG
LUSC
THCA
PRAD
SKCM
OV
BLCA
COAD
GBM
LIHC
KIRC
CESC
KIRP
SARC
ESCA
PCPG
PAAD
TGCT
READ
LAML
THYM
ACC
MESO
UVM
KICH
UCS
CHOL
DLBC


validation_dataset
0
20
40
60
80
100
120
140
160
168
98
89
86
85
85
82
80
80
73
71
69
65
62
56
51
48
41
33
31
30
26
25
20
20
17
15
13
12
11
9
7
BRCA
LUAD
UCEC
HNSC
LUSC
LGG
THCA
SKCM
PRAD
OV
BLCA
COAD
GBM
LIHC
KIRC
KIRP
CESC
SARC
ESCA
PCPG
PAAD
READ
TGCT
LAML
THYM
ACC
UVM
MESO
UCS
KICH
CHOL
DLBC

The detailed algorithm used for stratified sampling is shown below:

1. The stratified sampling

1. choose k for fold numbers.
2. use pandas data frame to group the data by cancer types

Cancer type	Sbs1	Sbs2	Gene1	Sample_id
ACC	32.12	45.76	1	0
ACC	34.12	0	1	1
ACC	33.12	40	1	2
…	…	…	1	..,
BLCA	56	98	0	50
BLCA	67	99	0	51
…	…	…	0	…
BRCA	23	34	1	100
…	…	…	1	…

3. for each cancer type and the data in that cancer type:

   reset index of each cancer type's data with indices as length of data in each cancer types

   Eg.

   ACC's indexes would be [0,1,2,3,⋯.49]
   BLCA's indexes would be [0,1,2,3,⋯,49]
   BRCA's indexes would be [0,1,2,3,⋯49]
   …

   We shuffle the index in each data of each cancer types

   Eg.

   ACC's indexes would now be [0,3,2,40,1,49⋯]
   BLCA's indexes would now be [0,2,1,48,36,⋯]
   BRCA's indexes would now be [0,32,1,18,⋯.]
   …

   Then ,we append those indexes in each cancers and their number of pieces to be put in each fold into indexe_list

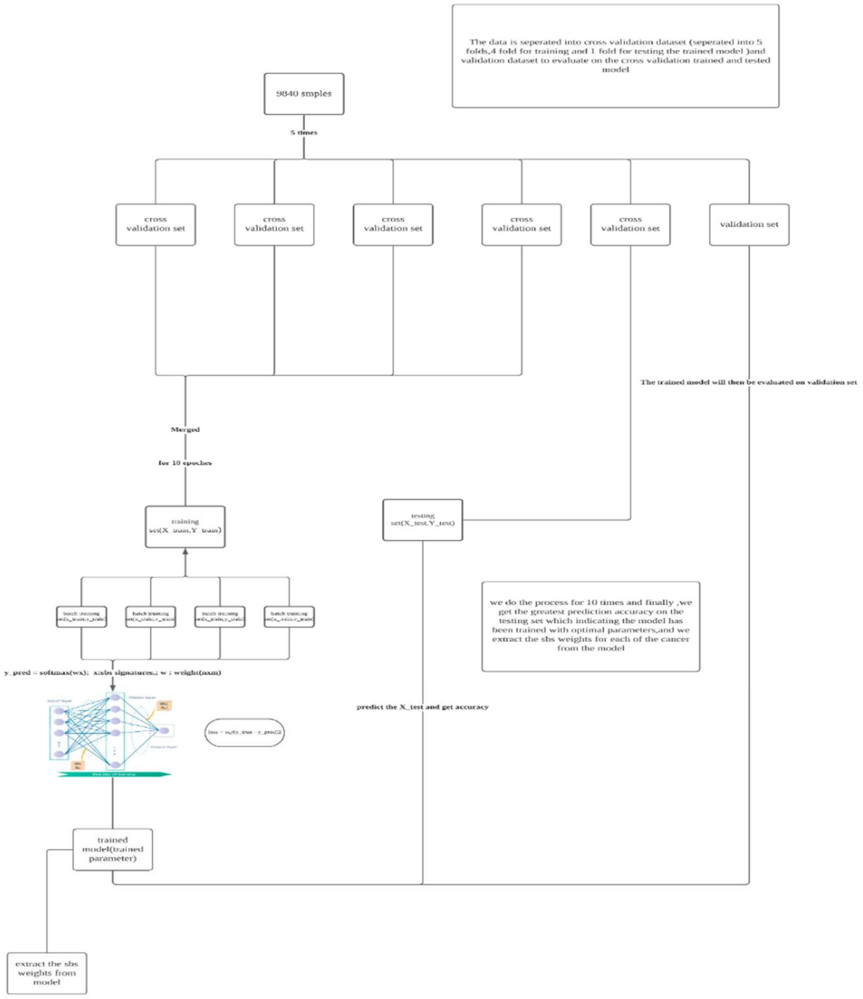
   Eg.

   The list contains:

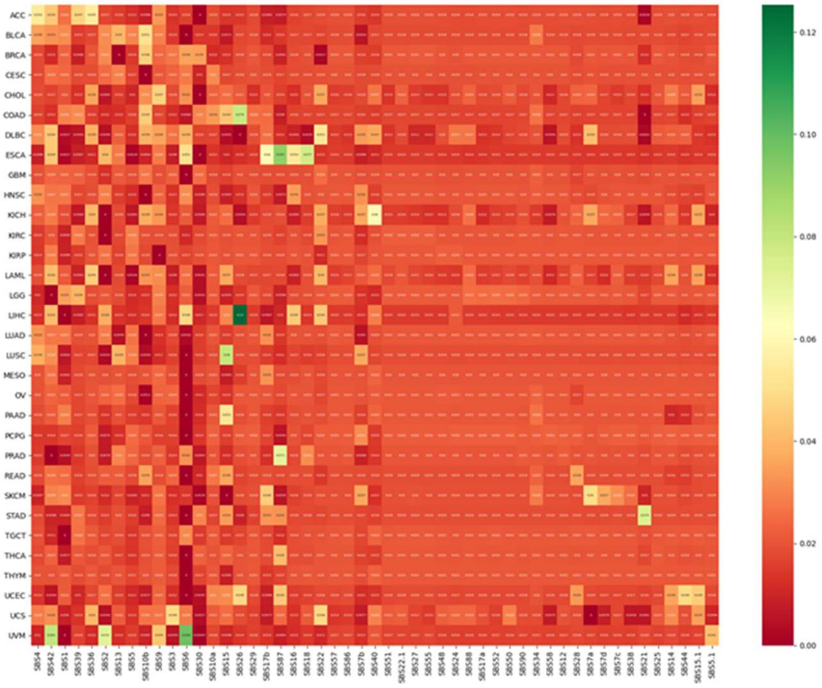
   ACC: ([0,3,2,40,1,49⋯] , 50/k)
   BLCA: ([0,2,1,48,36,⋯] ,50/k)
   BRCA: ([0,32,1,18,⋯.], 50/k)

The graph below shows the classification working flow for both models.

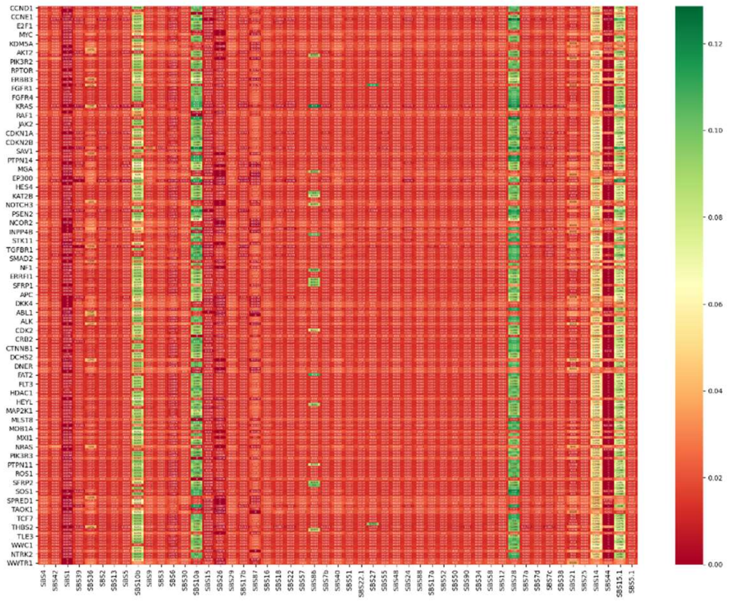
The approach is to separate the whole data into 6 datasets (5 cross validation dataset and 1 validation dataset) in each iteration of 5 iterations, we select the ith dataset in 5 cross validation datasets as testing set and rest as training set using the BPnet and at each iteration, when evaluate the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


The data is seperated into cross validation dataset (seperated into 5 folds,4 fold for training and 1 fold for testing the trained model)and validation dataset to evaluate on the cross validation trained and tested model
9840 smples
5 times
cross validation set
cross validation set
cross validation set
cross validation set
cross validation set
validation set
The trained model will then be evaluated on validation set
Merged
for 10 epoches
training set(X_train,Y_train)
testing set(X_test,Y_test)
batch training set(x_train,y_train)
batch training set(x_train,y_train)
batch training set(x_train,y_train)
batch training set(x_train,y_train)
we do the process for 10 times and finally ,we get the greatest prediction accuracy on the testing set which indicating the model has been trained with optimal parameters,and we extract the sbs weights for each of the cancer from the model
y_pred = softmax(wx); x:sbs signatures,; w : weight(nxm)
predict the X_test and get accuracy
loss = sum(y_true - y_pred)2
trained model(trained parameter)
extract the sbs weights from model

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

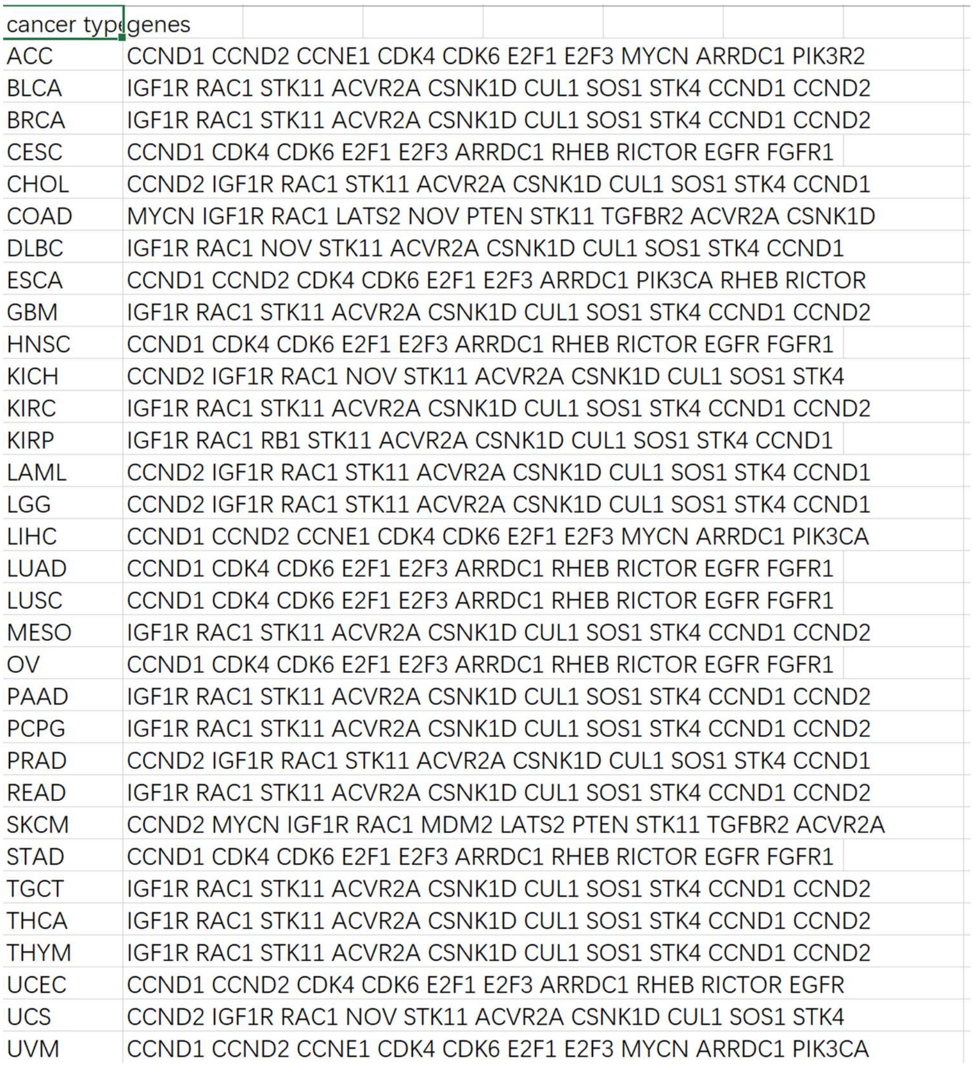


This is for the sbs weight in each gene.



(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. This is for the future for building the recommender system based on those top 10 genes



cancer typ	genes
ACC	CCND1 CCND2 CCNE1 CDK4 CDK6 E2F1 E2F3 MYCN ARRDC1 PIK3R2
BLCA	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
BRCA	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
CESC	CCND1 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR FGFR1
CHOL	CCND2 IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1
COAD	MYCN IGF1R RAC1 LATS2 NOV PTEN STK11 TGFBR2 ACVR2A CSNK1D
DLBC	IGF1R RAC1 NOV STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1
ESCA	CCND1 CCND2 CDK4 CDK6 E2F1 E2F3 ARRDC1 PIK3CA RHEB RICTOR
GBM	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
HNSC	CCND1 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR FGFR1
KICH	CCND2 IGF1R RAC1 NOV STK11 ACVR2A CSNK1D CUL1 SOS1 STK4
KIRC	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
KIRP	IGF1R RAC1 RB1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1
LAML	CCND2 IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1
LGG	CCND2 IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1
LIHC	CCND1 CCND2 CCNE1 CDK4 CDK6 E2F1 E2F3 MYCN ARRDC1 PIK3CA
LUAD	CCND1 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR FGFR1
LUSC	CCND1 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR FGFR1
MESO	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
OV	CCND1 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR FGFR1
PAAD	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
PCPG	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
PRAD	CCND2 IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1
READ	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
SKCM	CCND2 MYCN IGF1R RAC1 MDM2 LATS2 PTEN STK11 TGFBR2 ACVR2A
STAD	CCND1 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR FGFR1
TGCT	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
THCA	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
THYM	IGF1R RAC1 STK11 ACVR2A CSNK1D CUL1 SOS1 STK4 CCND1 CCND2
UCEC	CCND1 CCND2 CDK4 CDK6 E2F1 E2F3 ARRDC1 RHEB RICTOR EGFR
UCS	CCND2 IGF1R RAC1 NOV STK11 ACVR2A CSNK1D CUL1 SOS1 STK4
UVM	CCND1 CCND2 CCNE1 CDK4 CDK6 E2F1 E2F3 MYCN ARRDC1 PIK3CA

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 gene1 : 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).

Solved the class distance problem mentioned earlier by supervisor , I changed the loss calculation from finding error on the class label to finding the error from the vector representation

(5).

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. And worked on part of the classification method part.

2.The problems:

1. The gene of some of the patients in the cancer type may have mutated and the same gene may not be mutated in other patients in the same cancer type.so. the top 10 gene may not be indicative in this senario.

   The solution that I have thought about is to find the number of patients that have that specific gene mutated in their body and total number of patients have that gene recorded and we divide it to get the percentage of the patients that have the gene mutated and we sort the 187 genes based on this and strippe out lower occuraced genes to ensure those genes are frequently mutated in the patients of that cancers,then we calculate the weight of those genes in cancer and recommend those top 10 genes are the most related to this cancer.

3.The plan

(1). Solve the problem of weighted genes in each cancer.

(3). working on method part of the dissertation