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 classification is based on ['ACC', 'BLCA', 'BRCA', 'CESC', 'CHOL', 'COAD', 'DLBC', 'ESCA', 'GBM', 'HNSC', 'KICH', 'KIRC', 'KIRP', 'LAML', 'LGG', 'LIHC', 'LUAD', 'LUSC', 'MESO', 'OV', 'PAAD', 'PCPG', 'PRAD', 'READ', 'SKCM', 'STAD', 'TGCT', 'THCA', 'THYM', 'UCEC', 'UCS', 'UVM'] 32 of cancers

and ['CCND1', 'CCND2', 'CCND3', 'CCNE1', 'CDK4', 'CDK6', 'E2F1', 'E2F3', 'YAP1', 'MYC', 'MYCN', 'ARRDC1','KDM5A', 'NFE2L2', 'AKT1', 'AKT2', 'PIK3CA', 'PIK3CB', 'PIK3R2', 'RHEB', 'RICTOR', 'RPTOR', 'EGFR', 'ERBB2','ERBB3', 'PDGFRA', 'MET', 'FGFR1', 'FGFR2', 'FGFR3', 'FGFR4', 'KIT', 'IGF1R', 'KRAS', 'HRAS', 'BRAF', 'RAF1','RAC1', 'MAPK1', 'JAK2', 'MDM2', 'MDM4', 'CDKN1A', 'CDKN1B', 'CDKN2A', 'CDKN2B', 'CDKN2C', 'RB1', 'SAV1', 'LATS1','LATS2', 'PTPN14', 'NF2', 'FAT1', 'MGA', 'CNTN6', 'CREBBP', 'EP300', 'HES2', 'HES3', 'HES4', 'HES5', 'HEY1','KAT2B', 'NOTCH1', 'NOTCH2', 'NOTCH3', 'NOTCH4', 'NOV', 'PSEN2', 'FBXW7', 'NCOR1', 'NCOR2', 'KEAP1', 'CUL3','INPP4B', 'PIK3R1', 'PTEN', 'STK11', 'TSC1', 'TSC2', 'TGFBR1', 'TGFBR2', 'ACVR2A', 'SMAD2', 'SMAD3', 'SMAD4','NF1', 'RASA1', 'CBL', 'ERRFI1', 'TP53', 'ATM', 'SFRP1', 'ZNRF3', 'AMER1', 'APC', 'AXIN1', 'DKK1', 'DKK4', 'RNF43','TCF7L2', 'ABL1', 'ACVR1B', 'AKT3', 'ALK', 'ARAF', 'AXIN2', 'CDK2', 'CHEK2', 'CRB1', 'CRB2', 'CSNK1D', 'CSNK1E','CTNNB1', 'CUL1', 'DCHS1', 'DCHS2', 'DKK2', 'DKK3', 'DNER', 'ERBB4', 'ERF', 'FAT2', 'FAT3', 'FAT4', 'FLT3', 'GRB2','GSK3B', 'HDAC1', 'HES1', 'HEY2', 'HEYL', 'JAG2', 'MAML3', 'MAP2K1', 'MAP2K2', 'MAX', 'MLST8', 'MLX', 'MNT','MOB1A', 'MOB1B', 'MTOR', 'MXI1', 'NPRL2', 'NPRL3', 'NRAS', 'NTRK1', 'NTRK3', 'PIK3R3', 'PLXNB1', 'PPP2R1A','PTPN11', 'RET', 'RIT1', 'ROS1', 'RPS6KA3', 'RPS6KB1', 'SFRP2', 'SFRP4', 'SFRP5', 'SOS1', 'SOST', 'SPEN', 'SPRED1','STK3', 'STK4', 'TAOK1', 'TAOK2', 'TAOK3', 'TCF7', 'TCF7L1', 'TEAD2', 'THBS2', 'TLE1', 'TLE2', 'TLE3', 'TLE4','WIF1', 'WWC1', 'CRB3', 'LRP5', 'NTRK2', 'PDGFRB', 'TEAD3', 'WWTR1'] total 187 genes

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

[0.8640350877192983,0.8521303258145363,0.8602756892230576,0.8665413533834586, **0.8809523809523809**]

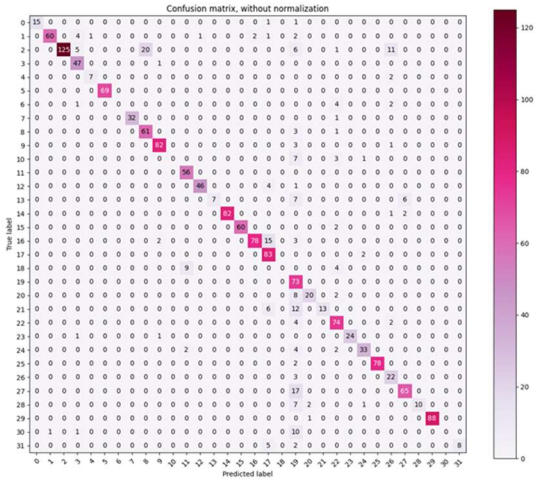
The validation accuracies for 5-fold cross validation are:

[0.8552472858866104, 0.8661037394451147, **0.8799758745476478**, 0.8395657418576599, 0.8600723763570567]

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	1.00	0.88	0.94	17
1	0.98	0.85	0.91	71
2	1.00	0.74	0.85	168
3	0.80	0.98	0.88	48
4	0.88	0.78	0.82	9
5	1.00	1.00	1.00	69
6	0.00	0.00	0.00	7
7	1.00	0.97	0.98	33
8	0.75	0.94	0.84	65
9	0.95	0.95	0.95	86
10	0.00	0.00	0.00	11
11	0.84	1.00	0.91	56
12	0.98	0.90	0.94	51
13	1.00	0.35	0.52	20
14	1.00	0.96	0.98	85
15	1.00	0.97	0.98	62
16	0.97	0.80	0.88	98
17	0.72	0.98	0.83	85
18	0.00	0.00	0.00	13
19	0.42	1.00	0.59	73
20	0.87	0.67	0.75	30
21	1.00	0.42	0.59	31
22	0.79	0.93	0.85	80
23	1.00	0.92	0.96	26
24	0.89	0.80	0.85	41
25	1.00	0.97	0.99	80
26	0.54	0.88	0.67	25
27	0.89	0.79	0.84	82
28	1.00	0.50	0.67	20
29	1.00	0.99	0.99	89
30	0.00	0.00	0.00	12
31	1.00	0.53	0.70	15
accuracy			0.86	1658
macro avg	0.79	0.73	0.74	1658
weighted avg	0.88	0.86	0.85	1658



As we can see, most of the 6th class(**'DLBC'**) samples were predicted to class 3(**'CESC'**)(1),class 22(**'PRAD'**)(4),class 26(**'TGCT'**)(2)

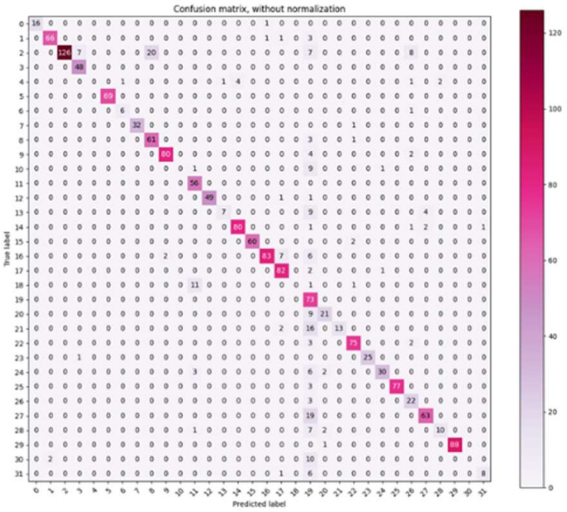
Most of the 10th class(**'KICH'**) samples were predicted to class 19 (**'PCPG'**) (7), class 22(**'PRAD') (**3) and class 24(**'STAD'**)(1)

Most of the 18th class(**'MESO'**) were predicted to class 11(**'KIRC'**) (9), class 22(**'PRAD'**)(4)

Most of the 30th class(**'UCS'**) were predicted to class 1(**'BLCA'**)(1),class 3(**'CESC'**)(1),class 19(**'OV'**)(10)

The second time: **87%,** The cancers have 0 accuracy are 10,18,30.

	precision	recall	f1-score	support
0	1.00	0.82	0.90	17
1	0.97	0.90	0.93	71
2	1.00	0.79	0.88	168
3	0.89	1.00	0.94	48
4	0.88	0.78	0.82	9
5	1.00	1.00	1.00	69
6	0.67	0.57	0.62	7
7	1.00	0.97	0.98	33
8	0.81	0.92	0.86	65
9	0.95	0.93	0.94	86
10	0.00	0.00	0.00	11
11	0.78	1.00	0.88	56
12	1.00	0.86	0.93	51
13	1.00	0.45	0.62	20
14	1.00	0.95	0.98	85
15	1.00	0.97	0.98	62
16	0.88	0.91	0.89	98
17	1.00	0.96	0.98	85
18	0.00	0.00	0.00	13
19	0.36	1.00	0.53	73
20	0.81	0.57	0.67	30
21	0.81	0.42	0.55	31
22	0.92	0.91	0.92	80
23	0.96	0.96	0.96	26
24	0.94	0.78	0.85	41
25	1.00	0.99	0.99	80
26	0.63	0.88	0.73	25
27	0.92	0.79	0.85	82
28	1.00	0.50	0.67	20
29	1.00	0.99	0.99	89
30	0.00	0.00	0.00	12
31	1.00	0.53	0.70	15
accuracy			0.87	1658
macro avg	0.82	0.75	0.77	1658
weighted avg	0.90	0.87	0.87	1658



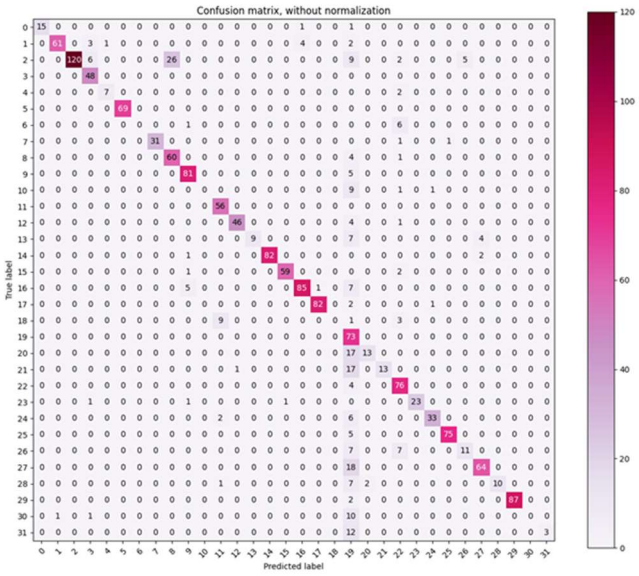
Most of the 10th class('KICH') samples were predicted to class 11 (**'KIRC'**) (1), class 19(**'OV'**) (9) and class 24(**'STAD'**)(1)

Most of the 18th class('MESO') were predicted to class 11(**'KIRC'**) (11), class 19(**'OV'**)(1) and class 22(**'PRAD'**)(1)

Most of the 30th class('UCS') were predicted to class 1(**'BLCA'**), class 19(**'OV'**)(10)

The third time: **88%**, The cancers have 0 accuracy are 10,18,30.

	precision	recall	f1-score	support
0	1.00	0.94	0.97	17
1	0.97	0.94	0.96	71
2	1.00	0.77	0.87	168
3	0.89	1.00	0.94	48
4	0.00	0.00	0.00	9
5	1.00	1.00	1.00	69
6	0.50	0.14	0.22	7
7	1.00	0.94	0.97	33
8	0.73	0.94	0.82	65
9	0.98	0.95	0.96	86
10	0.00	0.00	0.00	11
11	0.79	1.00	0.88	56
12	0.92	0.94	0.93	51
13	1.00	0.45	0.62	20
14	0.95	0.94	0.95	85
15	1.00	0.97	0.98	62
16	0.98	0.92	0.95	98
17	0.97	0.99	0.98	85
18	0.00	0.00	0.00	13
19	0.68	1.00	0.81	73
20	0.96	0.73	0.83	30
21	0.45	0.94	0.61	31
22	0.66	0.97	0.79	80
23	0.96	0.92	0.94	26
24	0.85	0.80	0.83	41
25	0.99	0.96	0.97	80
26	0.67	0.64	0.65	25
27	0.90	0.85	0.88	82
28	0.83	0.50	0.62	20
29	1.00	0.98	0.99	89
30	0.00	0.00	0.00	12
31	0.89	0.53	0.67	15
accuracy			0.88	1658
macro avg	0.77	0.74	0.74	1658
weighted avg	0.88	0.88	0.87	1658



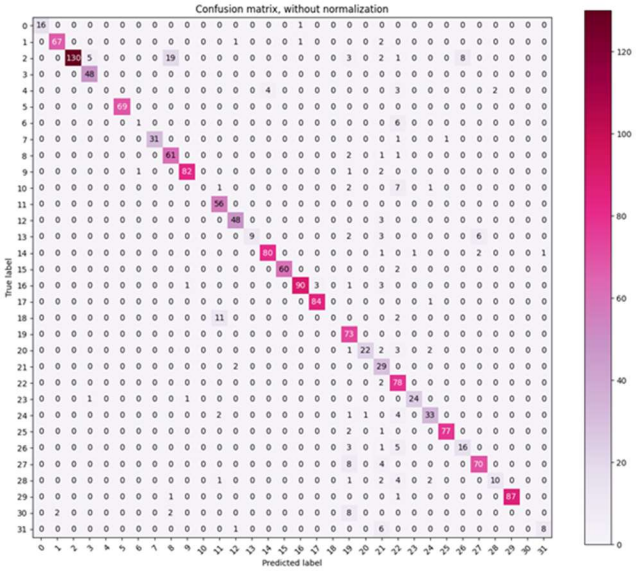
Most of the 10th class(**'KICH'**) samples were predicted to class 22 (**'PRAD'**) (1), class 19(**'OV'**) (9) and class 24(**'STAD'**)(1)

Most of the 18th class(**'MESO'**) were predicted to class 11(**'KIRC'**) (9), class 19(‘**OV’**)(1) and class 22(**'PRAD'**)(3)

Most of the 30th class(**'UCS'**) were predicted to class 1(**'BLCA'**)(1),class 3(**'CESC'**)(1) class 19(**'OV'**)(10)

The fourth time: **84%**, The cancers have 0 accuracy are 6,10,18,30.

	precision	recall	f1-score	support
0	1.00	0.88	0.94	17
1	0.98	0.86	0.92	71
2	1.00	0.71	0.83	168
3	0.81	1.00	0.90	48
4	0.88	0.78	0.82	9
5	1.00	1.00	1.00	69
6	0.00	0.00	0.00	7
7	1.00	0.94	0.97	33
8	0.70	0.92	0.79	65
9	0.90	0.94	0.92	86
10	0.00	0.00	0.00	11
11	0.82	1.00	0.90	56
12	0.98	0.90	0.94	51
13	1.00	0.45	0.62	20
14	1.00	0.96	0.98	85
15	0.98	0.95	0.97	62
16	0.94	0.87	0.90	98
17	0.99	0.96	0.98	85
18	0.00	0.00	0.00	13
19	0.32	1.00	0.48	73
20	0.87	0.43	0.58	30
21	1.00	0.42	0.59	31
22	0.75	0.95	0.84	80
23	1.00	0.88	0.94	26
24	0.94	0.80	0.87	41
25	0.99	0.94	0.96	80
26	0.69	0.44	0.54	25
27	0.91	0.78	0.84	82
28	1.00	0.50	0.67	20
29	1.00	0.98	0.99	89
30	0.00	0.00	0.00	12
31	1.00	0.20	0.33	15
accuracy			0.84	1658
macro avg	0.80	0.70	0.72	1658
weighted avg	0.88	0.84	0.84	1658



As we can see, most of the 6th class(**'DLBC'**) samples were predicted to class 22(**'PRAD'**)(6)

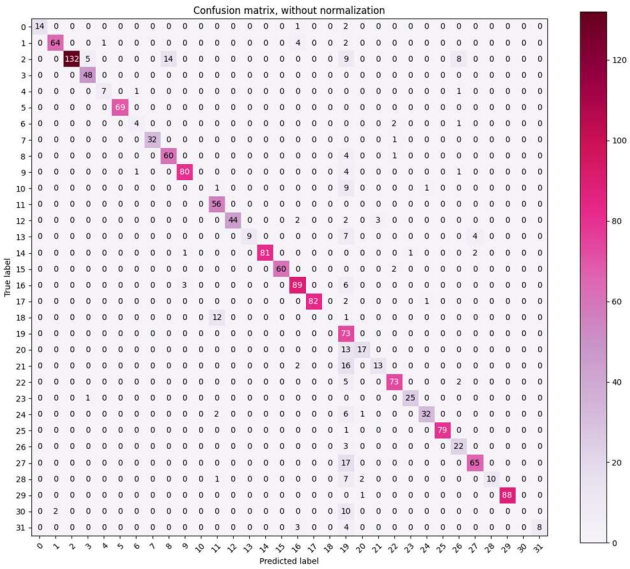
Most of the 10th class(**'KICH'**) samples were predicted to class 11 ('KIRC') (1), class 19(**'OV'**) (2), class 22(**'PRAD'**)(7),and class 24(**'STAD'**)(1)

Most of the 18th class(**'MESO'**) were predicted to class 11(**'KIRC'**) (11), class 22(**'PRAD'**)(2)

Most of the 30th class(**'UCS'**) were predicted to class 1(**'BLCA'**)(2),class 8(**'GBM'**)(2),class 19(**'OV'**)(8)

The fifth time: **86%**, The cancers have 0 accuracy are 10,18, 30.

	precision	recall	f1-score	support
0	1.00	0.94	0.97	17
1	0.97	0.93	0.95	71
2	1.00	0.75	0.86	168
3	0.86	1.00	0.92	48
4	0.00	0.00	0.00	9
5	1.00	1.00	1.00	69
6	0.86	0.86	0.86	7
7	1.00	0.97	0.98	33
8	0.75	0.94	0.84	65
9	0.98	0.93	0.95	86
10	0.00	0.00	0.00	11
11	0.78	1.00	0.88	56
12	1.00	0.96	0.98	51
13	0.88	0.35	0.50	20
14	0.95	0.94	0.95	85
15	1.00	0.97	0.98	62
16	0.98	0.85	0.91	98
17	0.87	0.96	0.92	85
18	0.00	0.00	0.00	13
19	0.36	1.00	0.53	73
20	0.81	0.70	0.75	30
21	1.00	0.42	0.59	31
22	0.94	0.94	0.94	80
23	1.00	0.96	0.98	26
24	0.94	0.73	0.82	41
25	1.00	0.96	0.98	80
26	0.59	0.88	0.71	25
27	0.91	0.77	0.83	82
28	0.83	0.50	0.62	20
29	1.00	0.99	0.99	89
30	0.00	0.00	0.00	12
31	0.89	0.53	0.67	15
accuracy			0.86	1658
macro avg	0.79	0.74	0.75	1658
weighted avg	0.89	0.86	0.86	1658



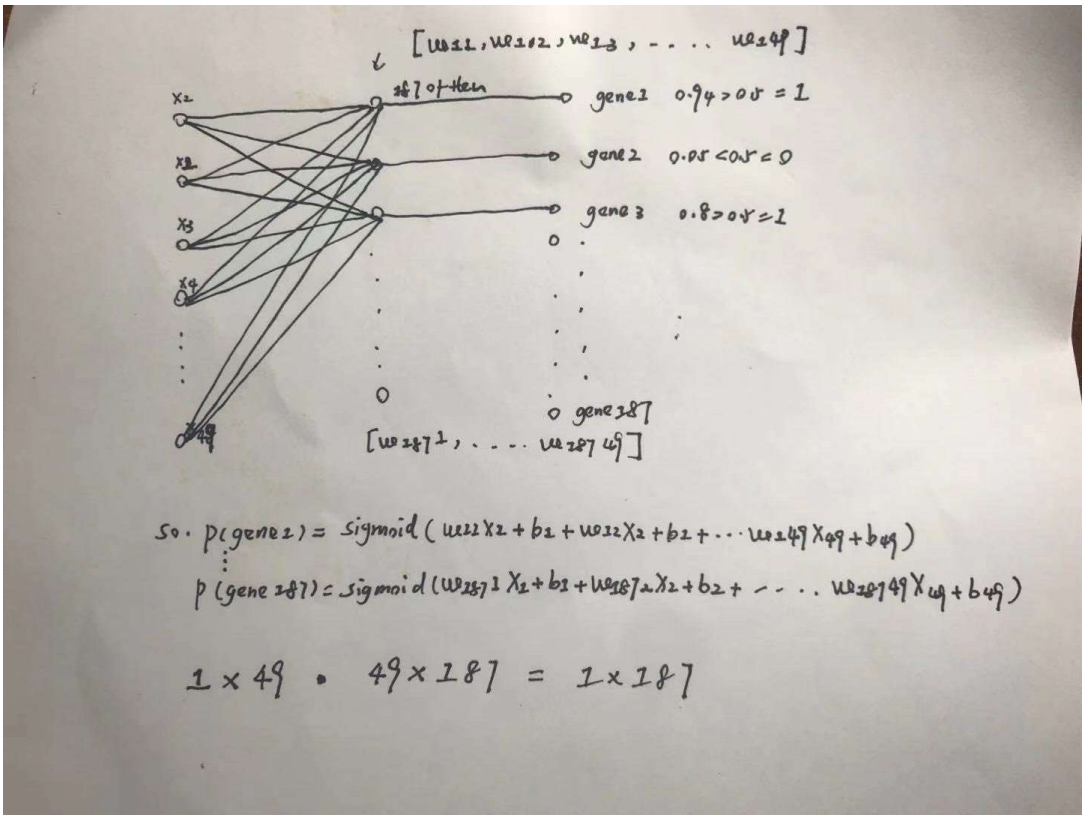
Most of the 10th class(**'KICH'**) samples were predicted to class 11 (**'KIRC'**) (1), class 19(**'OV'**) (9) and class 24(**'STAD'**)(1)

Most of the 18th class(**'MESO'**) were predicted to class 11(**'KIRC'**) (12), class 19(‘**OV'**)(1)

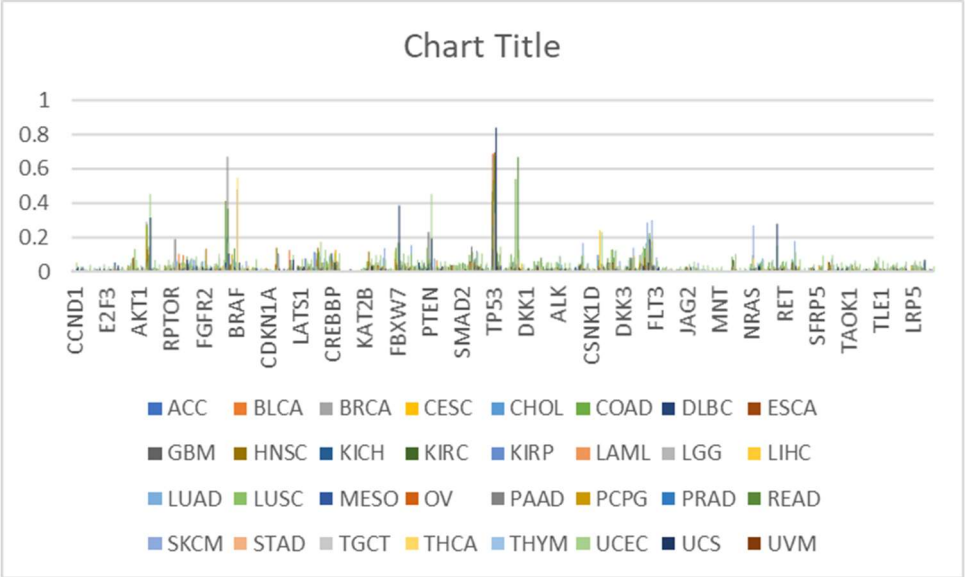
Most of the 30th class(**'UCS'**) were predicted to class 1(**'BLCA'**)(2), class 19(**'OV'**)(10)

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

The 5-fold cross validation has 5 testing result, they are:
[0.9943609022556391, 0.9949874686716792, 0.9943609022556391, 0.9949874686716792, **0.9962406015037594**]
The validation accuracies for 5-fold cross validation are:
[0.9933655006031363, 0.9939686369119421, 0.9933655006031363, **0.9981905910735827**, 0.9951749095295537]



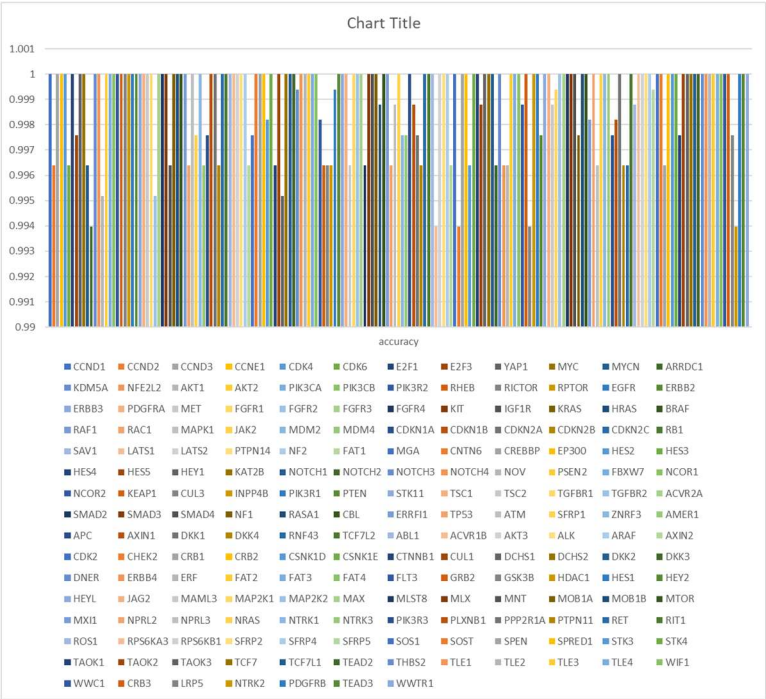
The graph shows the frequently mutated gene in each of the cancers.



Like the TP53 is 0.84 (**84% of the samples in UCS have TP53 mutated**) mutated in UCS samples, looks like most of the cancers has TP53 mutated.

The classification status in each fold across genes

Fold 1 for validation data



Fold 2 for validation data

Fold 3 for validation data

Fold 4 for validation data

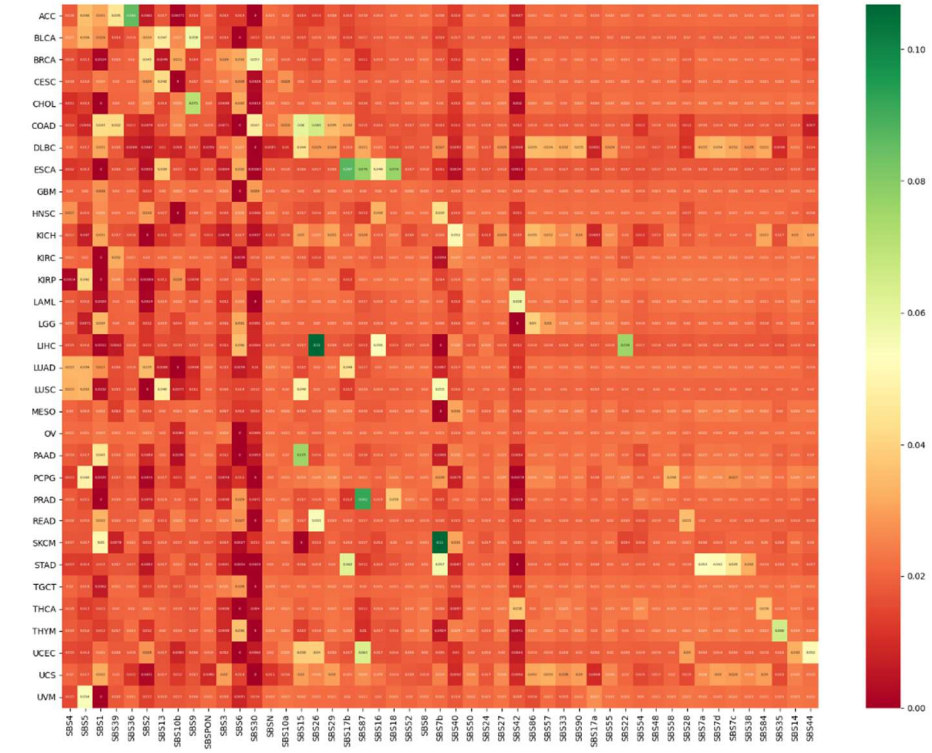
Fold 5 for validation data

(3). Find the most determinable genes in each cancer types by eliminated the gene

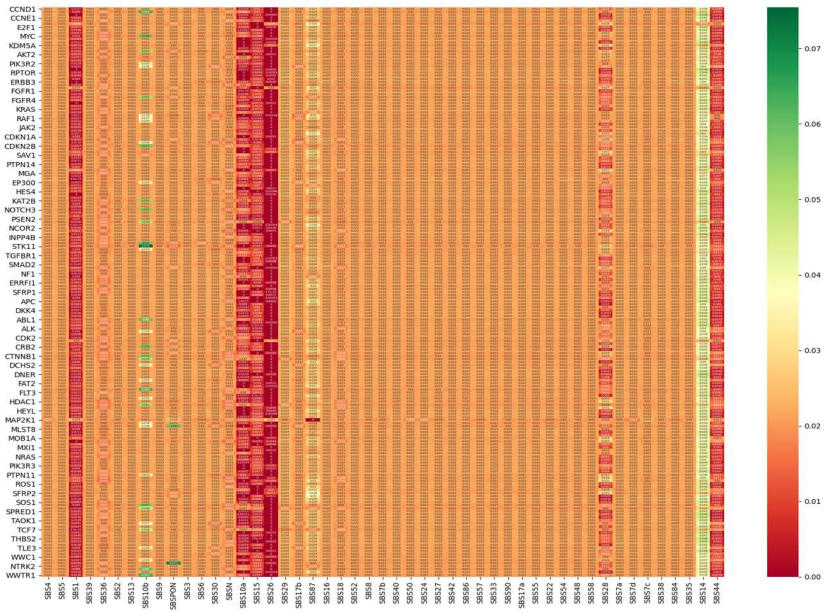
occurrence lesser than 5% and sbs intersection with that cancer greater than 0:

cancer typ	genes
ACC	CHEK2 TP53 NOTCH4 FAT1 RB1 FAT4 CTNNB1 ATM NF1 NOTCH3
BLCA	TP53 NOTCH4 FAT1 RB1 HRAS MET PIK3CA SPEN MTOR FAT4
BRCA	TP53 NF1 FAT1 PIK3CA AKT1 FAT3 DCHS2 PTEN ERBB2 NCOR1
CESC	WWC1 TP53 TSC2 NOTCH4 FAT1 RB1 JAK2 PIK3CA SPEN MTOR
CHOL	SPEN RIT1 NTRK1 FLT3 CHEK2 TP53 RASA1 NF1 TSC1 PIK3R1
COAD	WWC1 MAP2K1 DKK2 TP53 TSC2 INPP4B NOTCH4 FAT1 RB1 JAK2
DLBC	FAT1 CHEK2 TP53 INPP4B MET SPEN FAT4 CUL1 CRB1 ATM
ESCA	NTRK1 MTOR FLT3 FAT4 CTNNB1 CRB1 ALK ATM TP53 RASA1
GBM	TP53 RB1 PIK3CA FAT4 NF1 PIK3R1 PDGFRA PTEN NOTCH2 EGFR
HNSC	TP53 FAT1 HRAS PIK3CA SPEN FAT4 CRB1 RASA1 FBXW7 NOTCH3
KICH	TP53 TSC2 MTOR ATM PTEN
KIRC	SPEN TP53 FAT1 MTOR ATM ROS1 ERBB4
KIRP	SPEN FAT1 MET CREBBP CUL3
LAML	SPEN FLT3 TP53 PTPN11 KRAS
LGG	TP53 PIK3CA NF1 PTEN EGFR FAT2 NOTCH1
LIHC	FAT4 CTNNB1 TP53 NF1 NOTCH4 NOTCH3 RB1 IGF1R PIK3CA ATM
LUAD	TP53 INPP4B NOTCH4 FAT1 RB1 MET PIK3CA NTRK2 SPEN MTOR
LUSC	NTRK2 WIF1 SPEN NTRK1 FLT3 FAT4 CRB1 TP53 RASA1 NF1
MESO	TP53 FAT4 MGA LATS2 LATS1 NF2 PTEN NCOR1
OV	TP53 NOTCH4 FAT1 SPEN FAT4 ATM NF1 FAT3
PAAD	TP53 PIK3CA FAT4 CTNNB1 ATM TGFBR1 FAT3 SMAD4 EP300 TGFBR2
PCPG	NF1 HRAS RET
PRAD	TP53 ATM DCHS2
READ	WWC1 MAP2K1 TP53 INPP4B NOTCH4 FAT1 RB1 JAK2 PIK3CA SPEN
SKCM	MET WWC1 SFRP2 MTOR MAP2K1 DKK2 ALK ATM TP53 TSC2
STAD	WWC1 TP53 TSC2 NOTCH4 FAT1 RB1 JAK2 IGF1R PIK3CA NTRK2
TGCT	PIK3CA NRAS KIT KRAS
THCA	HRAS NRAS BRAF
THYM	TP53 HRAS NRAS
UCEC	WWC1 RPS6KB1 NPRL3 MLST8 MAP2K1 DKK2 CHEK2 TP53 TSC2 INPP4B
UCS	CHEK2 TP53 SAV1 RB1 MET PIK3CA NTRK1 FAT4 ATM NF1
UVM	RNF43 APC

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



This is for the sbs weight in each gene.



(4).

Tried transfer learning and it does not improve the accuracy of cancer classification but improved gene classification (make all validation set accuracy in each fold 100%) if I use the pre-defined model as the model that has trained on gene classification. And it would be worse if it's opposite. The reason why the accuracy of gene prediction is increased is because the number of genes to be predicted are limited to 32 gene.

(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 ethical permission
2.Data is too big to put on github,but the checker will need to run the code,how do I include the data

3.The plan

(3). working on the dissertation