Cancer Cell Cluster Detection by Unified Tumor Tissue Slide Image and Gene Expression Quantification

Wenkang Zhan, Zhengkang Fan
Dept. of Computer and Information Science

Background

Diagnosis of cancer region in a slide image is difficult and time-consuming, as cancer may transfer to other place.

Machine learning have many successful applications on image data. We want to use techniques to detect cancer cluster region.

In real scenarios, slide image data set is small. Gene expression quantification is available and it is a powerful material to classify tumors. Thus, we unify gene expression quantification and slide image to improve the performance of cancer cell cluster detection.

Data Collection

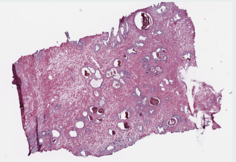
Data source: The Cancer Genome Atlas (TCGA)

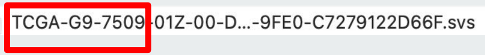
- Public-open database
- Provide over 20,000 cancers and matches 33 cancer types.
- https://portal.gdc.cancer.gov/

Data:

- Slide images
- Gene expression quantification -> feature
- Clinical data: diagnosis record -> label
- Meta data: Identity information ->tag for combining data
- supplemental clinical data: unprocessed data containing some useful information

Data Collection and Pre-processing

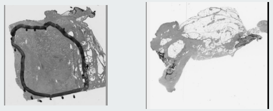




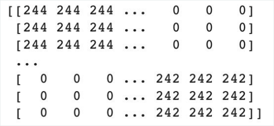
TCGA-G9-7509-01Z-00-D...-9FE0-C7279122D66F.svs

Format transformation:
svs->png; compression

Exact ID from file name



crop, rescale and to grayscale

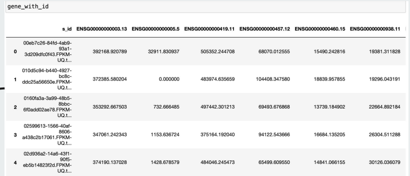


```
[[244 244 244 ...   0   0   0]
 [244 244 244 ...   0   0   0]
 [244 244 244 ...   0   0   0]
 ...
 [  0   0   0 ... 242 242 242]
 [  0   0   0 ... 242 242 242]
 [  0   0   0 ... 242 242 242]]
```

Image to numpy array

Pair-wise matching

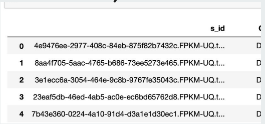
Pair-wise gene and image
data in matrix format
(training data)



gene_with_id

	s_id	ENSG00000000003.13	ENSG00000000005.5	ENSG00000000419.11	ENSG00000000457.12	ENSG00000000460.15	ENSG00000000938.11
0	00eb7c26-84fd-4ab9-93a1-3d209dfc0f43.FPKM-UQ.t...	392168.920789	32911.830937	505352.244708	68070.012555	15490.242816	19381.311828
1	010d5c94-b440-4927-bc8c-ddc25a56650e.FPKM-UQ.t...	372385.580204	0.000000	483974.635659	104408.347580	18839.957855	19296.043191
2	0160fa3a-3a99-48b5-8bbc-6f0add02ae78.FPKM-UQ.t...	353292.667503	732.666485	497442.301213	69493.676868	13739.184902	22664.892184
3	02599613-1566-40af-8606-a438c2b17061.FPKM-UQ.t...	347061.242343	1153.636724	375164.192040	94122.543666	16684.135205	26304.511288
4	02d936a2-14a6-43f1-90f5-eb5b14823f2d.FPKM-UQ.t...	374190.137028	1428.678579	484046.245473	65499.609550	14841.066155	30126.036079

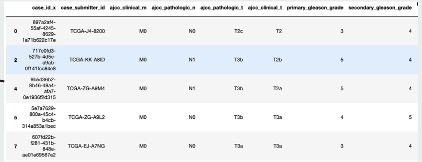
Gene expression quantification -> feature



	s_id	
0	4e9476ee-2977-408c-84eb-875f82b7432c.FPKM-UQ.t...	D
1	8aa4f705-5aac-4765-b686-73ee5273e465.FPKM-UQ.t...	D
2	3e1ecc6a-3054-464e-9c8b-9767fe35043c.FPKM-UQ.t...	D
3	23eaf5db-46ed-4ab5-ac0e-ec6bd65762d8.FPKM-UQ.t...	D
4	7b43e360-0224-4a10-91d4-d3a1e1d30ec1.FPKM-UQ.t...	D

Meta data: Identity information ->tag for
combining data

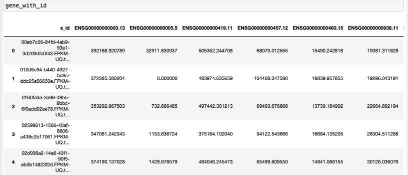
Generate pathological
stage (label)



	case_id_x	case_submitter_id	ajcc_clinical_m	ajcc_pathologic_n	ajcc_pathologic_t	ajcc_clinical_t	primary_gleason_grade	secondary_gleason_grade
0	897a2af4-55af-4245-8629-1a71b622c17e	TCGA-J4-8200	M0	N0	T2c	T2	3	4
2	717c0fd3-527b-4d5e-a9ab-0f141fcc84e8	TCGA-KK-A8ID	M0	N1	T3b	T2b	5	4
4	9b5d36b2-8b46-48a4-afa7-0e1936f2d315	TCGA-ZG-A9M4	M0	N1	T3b	T2a	5	4
5	5e7a7629-800a-45c4-b4cb-314a853a1bec	TCGA-ZG-A9L2	M0	N0	T3b	T3a	4	5
7	607fd22b-f281-431b-848e-ae01e69567e2	TCGA-EJ-A7NG	M0	N0	T3a	T3a	3	4

Clinical data: diagnosis record -> label

Feature Reduction - gene expression quantification

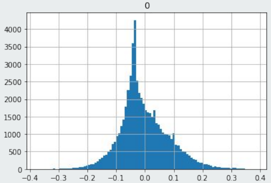


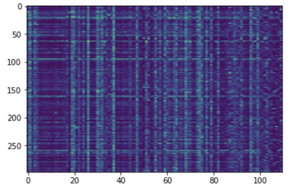
gene_with_id

	s_id	ENSG00000000003.13	ENSG00000000005.5	ENSG00000000419.11	ENSG00000000457.12	ENSG00000000460.15	ENSG00000000938.11
0	00eb7c26-84fd-4ab9-93a1-3d209dfc0f43.FPKM-UQ.t...	392168.920789	32911.830937	505352.244708	68070.012555	15490.242816	19381.311828
1	010d5c94-b440-4927-bc8c-ddc25a56650e.FPKM-UQ.t...	372385.580204	0.000000	483974.635659	104408.347580	18839.957855	19296.043191
2	0160fa3a-3a99-48b5-8bbc-6f0add02ae78.FPKM-UQ.t...	353292.667503	732.666485	497442.301213	69493.676868	13739.184902	22664.892184
3	02599613-1566-40af-8606-a438c2b17061.FPKM-UQ.t...	347061.242343	1153.636724	375164.192040	94122.543666	16684.135205	26304.511288
4	02d936a2-14a6-43f1-90f5-eb5b14823f2d.FPKM-UQ.t...	374190.137028	1428.678579	484046.245473	65499.609550	14841.066155	30126.036079

Gene expression quantification
(over 60,000 features)

Correlation of feature v.s. label



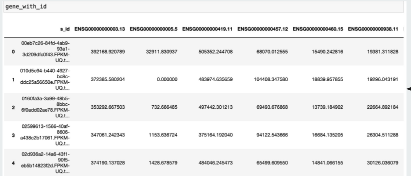


Set threshold to select features

Manually select meaningful features

Performance comparison with different subset of features by a baseline models

Gene expression quantification
(hundreds of features)



gene_with_id

	s_id	ENSG00000000003.13	ENSG00000000005.5	ENSG00000000419.11	ENSG00000000457.12	ENSG00000000460.15	ENSG00000000938.11
0	00eb7c26-84fd-4ab9-93a1-3d209dfc0f43.FPKM-UQ.t...	392168.920789	32911.830937	505352.244708	68070.012555	15490.242816	19381.311828
1	010d5c94-b440-4927-bc8c-ddc25a56650e.FPKM-UQ.t...	372385.580204	0.000000	483974.635659	104408.347580	18839.957855	19296.043191
2	0160fa3a-3a99-48b5-8bbc-6f0add02ae78.FPKM-UQ.t...	353292.667503	732.666485	497442.301213	69493.676868	13739.184902	22664.892184
3	02599613-1566-40af-8606-a438c2b17061.FPKM-UQ.t...	347061.242343	1153.636724	375164.192040	94122.543666	16684.135205	26304.511288
4	02d936a2-14a6-43f1-90f5-eb5b14823f2d.FPKM-UQ.t...	374190.137028	1428.678579	484046.245473	65499.609550	14841.066155	30126.036079

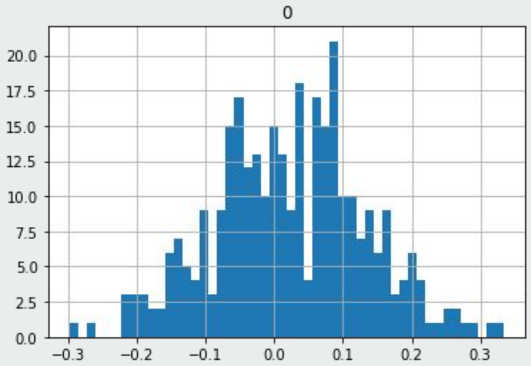
Feature Reduction - gene expression quantification

Number of remaining features with different thresholds

threshold	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.30
#features	3566	2333	1534	969	623	397	238	132

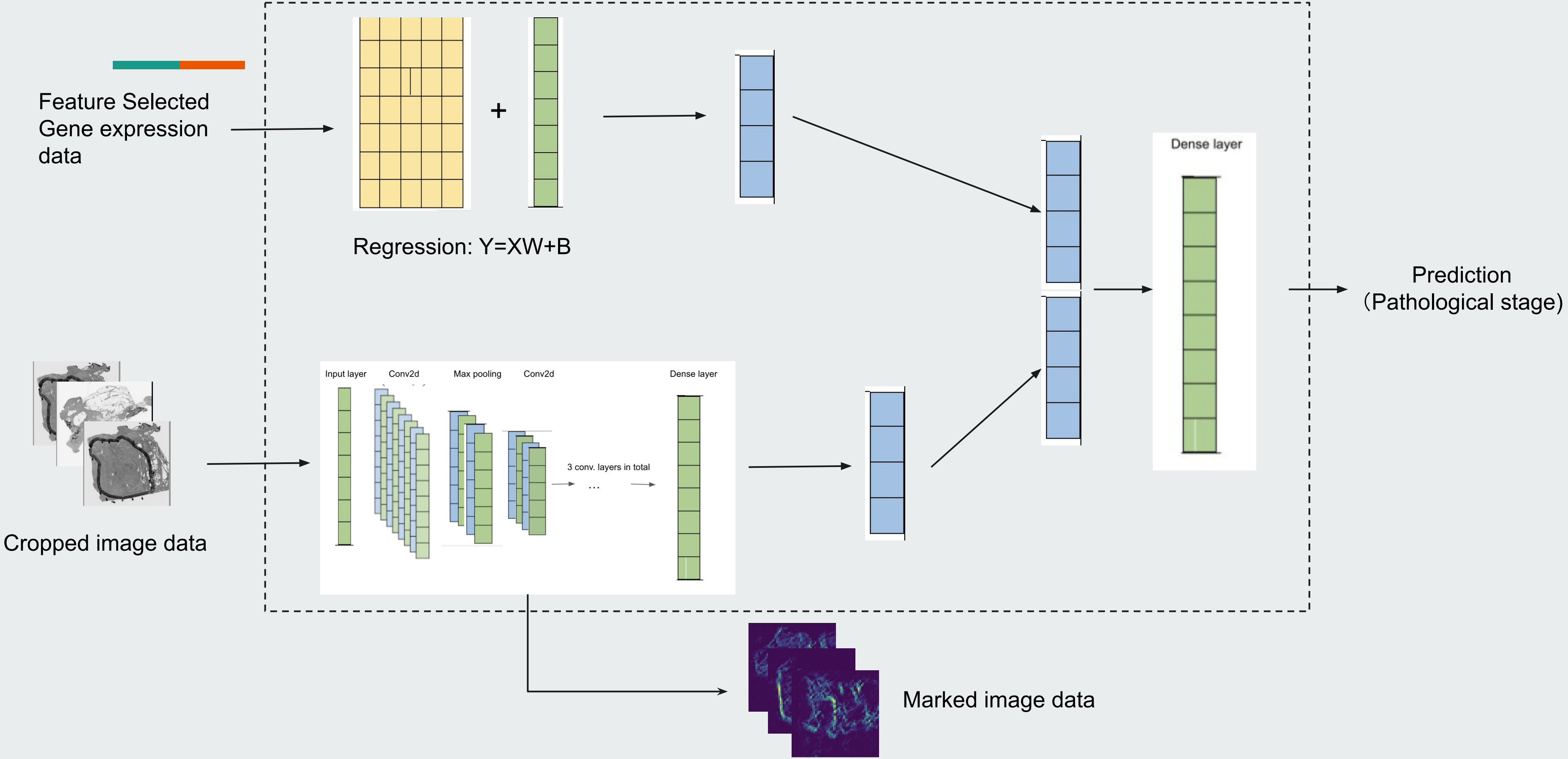
Accuracy of baselines - only on gene expression quantification

	Selected features by correlations								Domain knowledge
thres.	0.16	0.18	0.20	0.22	0.24	0.26	0.28	0.30	-
#feat.	3566	2333	1534	969	623	397	238	132	315
LR	**0.66**	0.66	0.6	0.63	0.63	0.6	0.56	0.56	**0.66**
CNN	0.44	0.43	0.43	0.48	0.48	**0.56**	**0.56**	**0.56**	**0.56**
LSTM	0.46	0.48	0.48	0.48	0.48	0.53	**0.56**	**0.56**	**0.56**



Correlation of feat. v.s. Label
(domain knowledge)

Model



Feature Selected Gene expression data

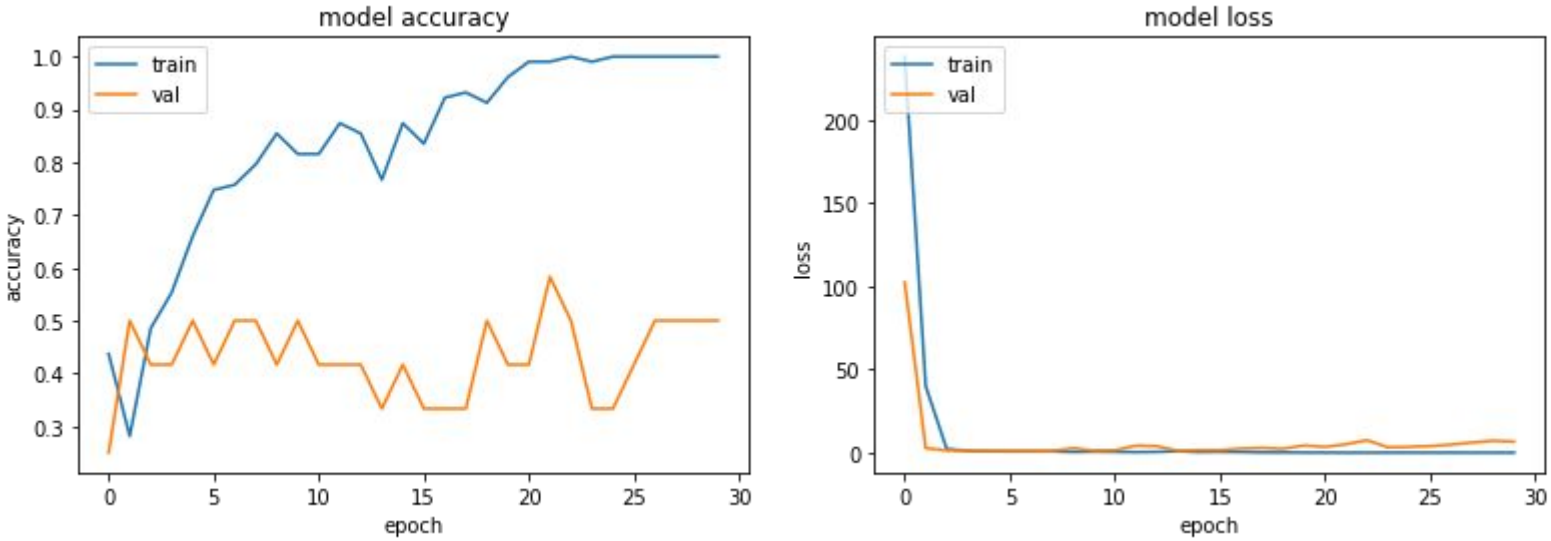
Regression: Y=XW+B

Cropped image data

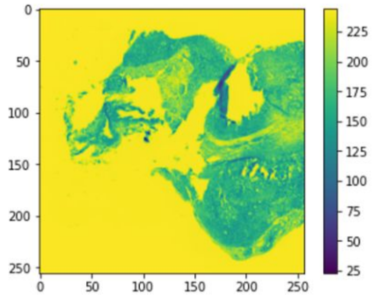
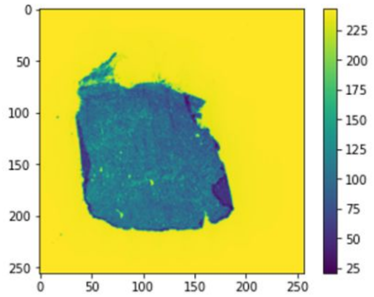
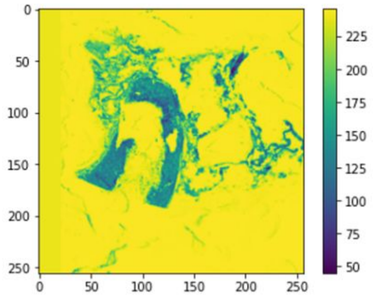
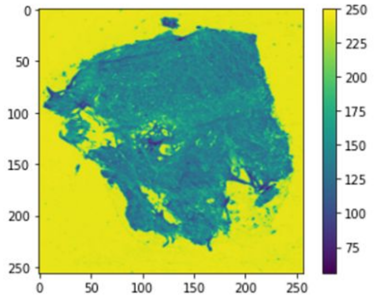
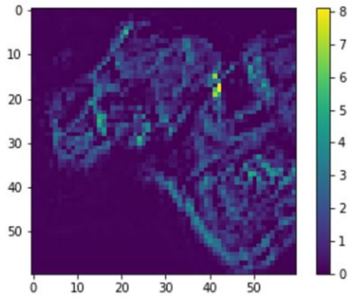
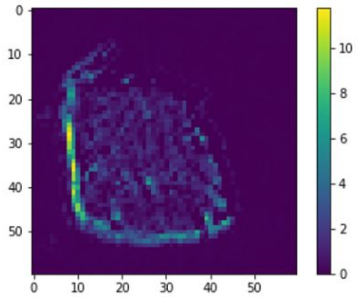
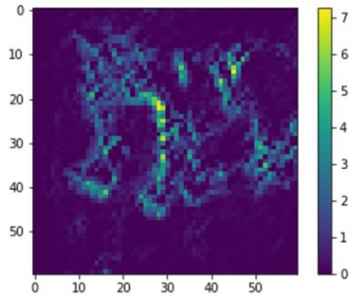
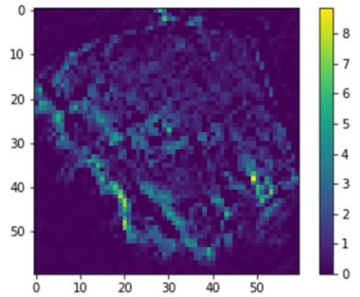
Prediction
（Pathological stage)

Marked image data

Results


model accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
accuracy
train
val
0
5
10
15
20
25
30
epoch
model loss
200
150
100
50
0
loss
train
val
0
5
10
15
20
25
30
epoch

Results

	1	2	3	4
Replot of Original Image				
Markers by outputs of 3rd conv layer				

Future

- Add more data into dataset(115 samples)

- Overfitting

- Improve the model structure

Thanks for your listening!

Q&A

Wenkang Zhan, Zhengkang Fan
Dept. of Computer and Information Science