Deep Learning Based Tumor Type Classification Using Gene Expression Data

Boyu Lyu
Virginia Tech

Anamul Haque
Virginia Tech

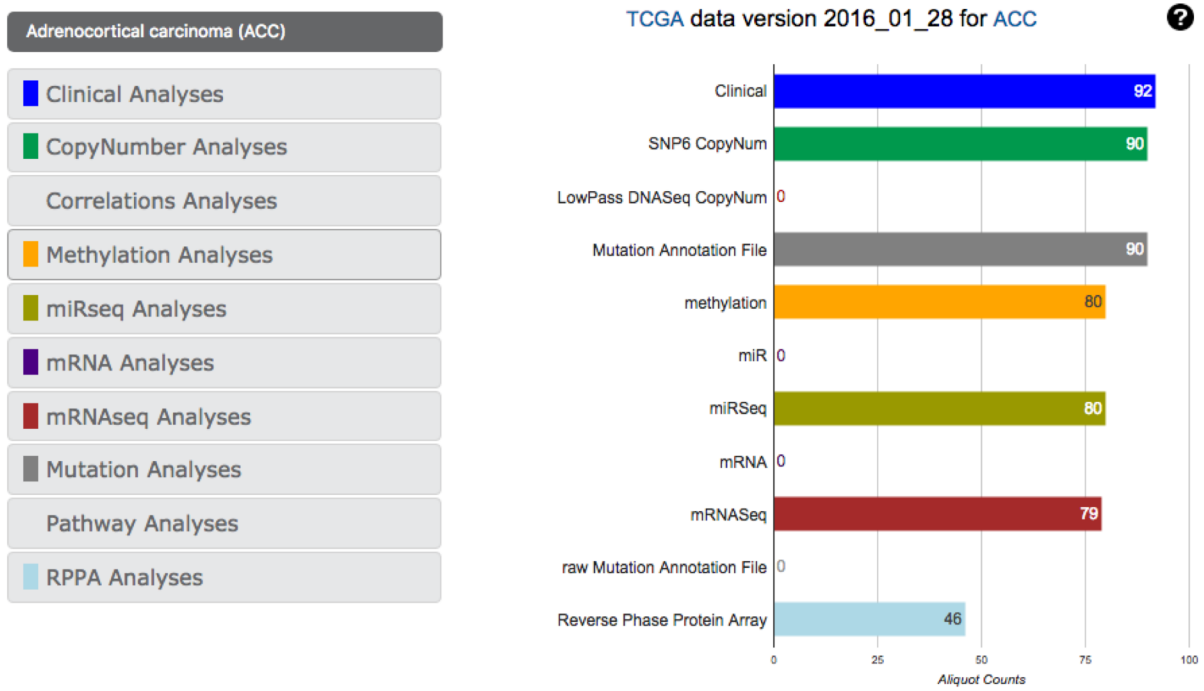
TCGA & Pan-Cancer Atlas

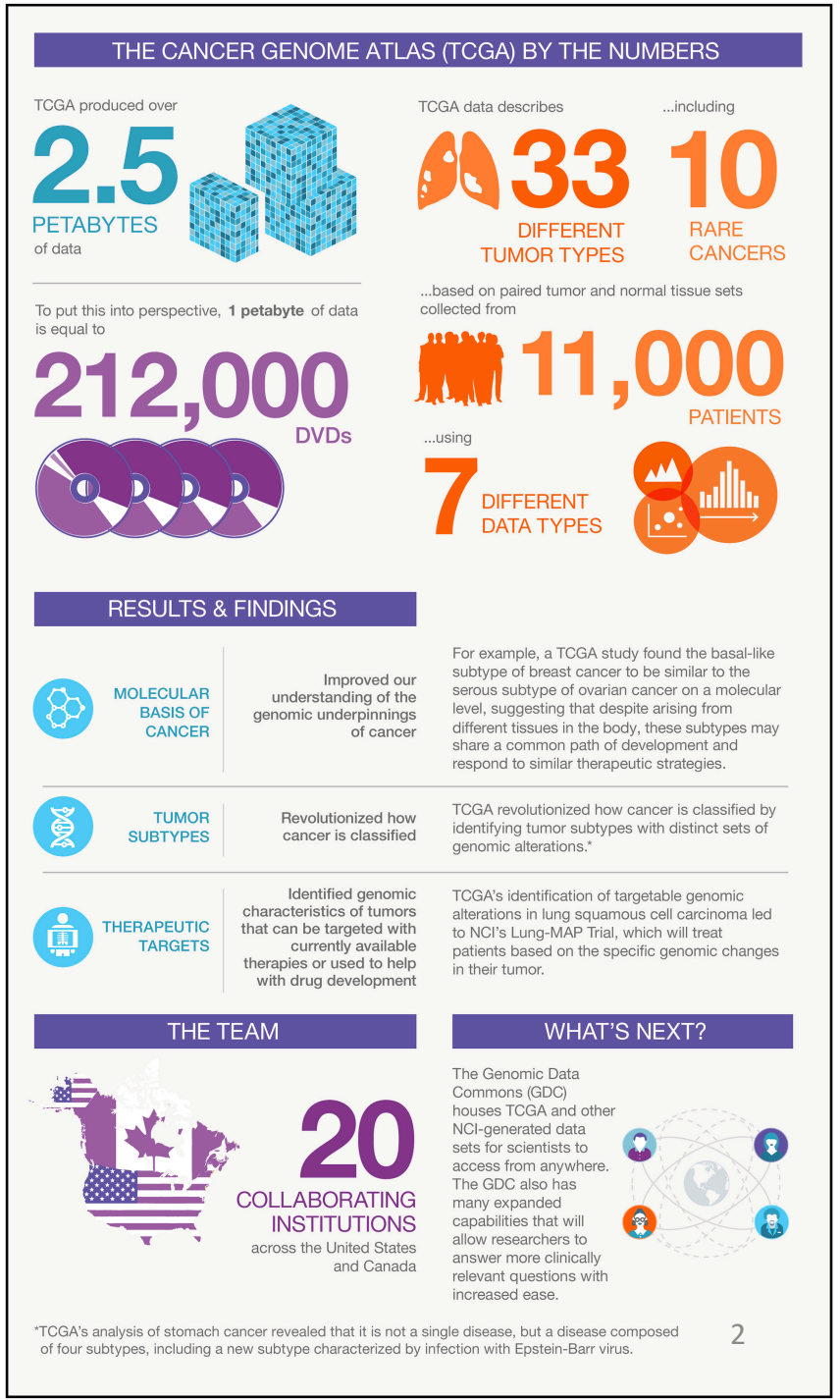
- Since Next Generation Sequencing (NGS) has been invented, large volume of genomic sequencing data has been collected. TCGA (https://cancergenome.nih.gov/) provides a huge knowledge source for the understanding of the cause of human tumors.

- With these tumor samples, since 2012, a project called “Pan-cancer” was launched to “assemble coherent, consistent TCGA data sets across tumor types, as well as across platforms, and then to analyze and interpret these data ”.

TCGA & Pan-Cancer Atlas

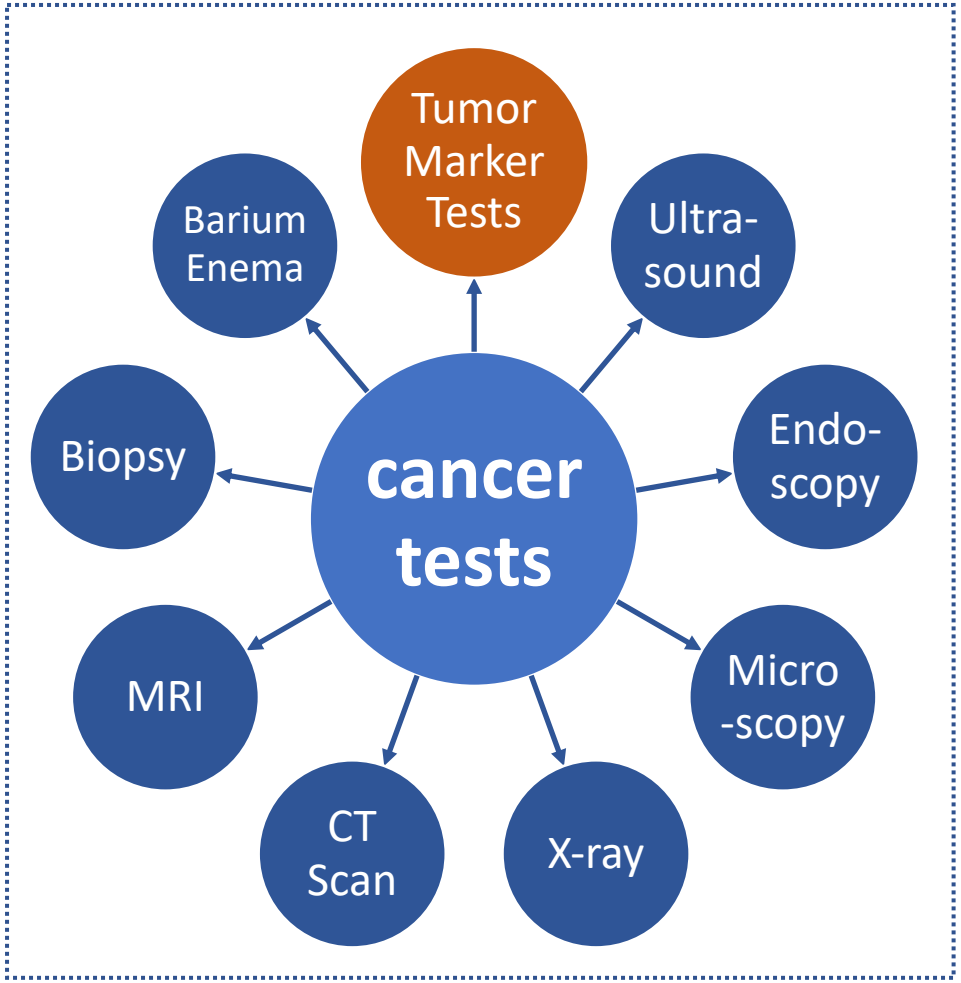
- Pan-Cancer Atlas (http://gdac.broadinstitute.org/), it contains the analysis of **11,000** tumors from **33** of the most prevalent cancers.





These analysis results include RNA-Seq, copy number, mutation annotation, methylation and detailed clinical records. Which can be used for further exploration. **Now, this project is coming to an end. But the study on this dataset is still at the beginning.**

Tumor type classification using the genome --- earlier diagnosis



https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures

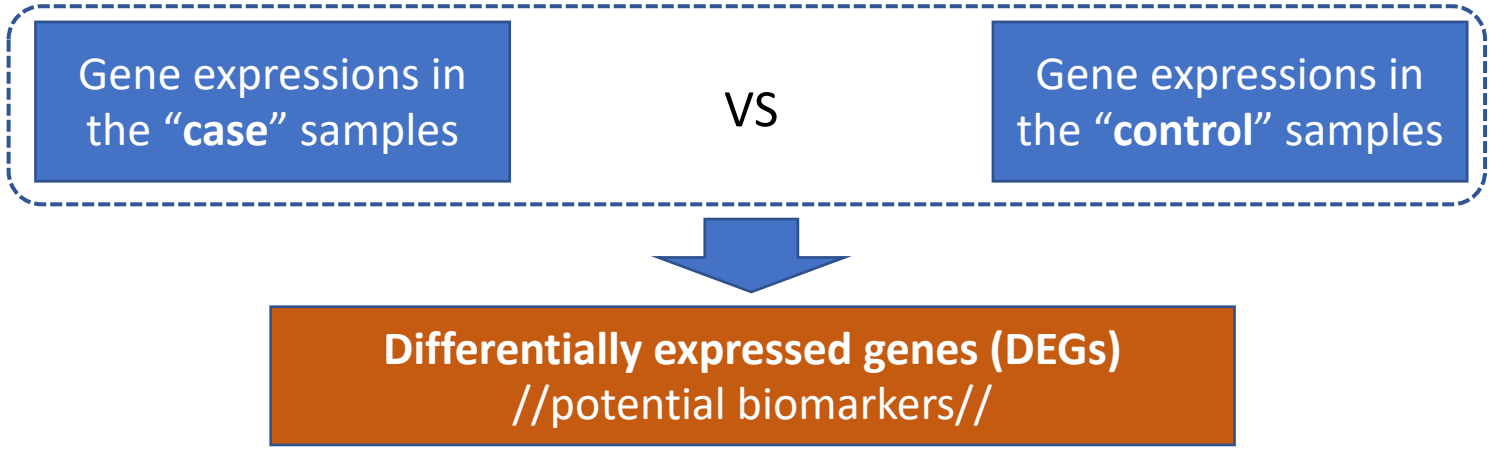
- Most tests are based on the visualization of the change in tissues or cells.
- Genome variation is one step earlier than the morphological change.

Genome tests focusing on the change of the gene expression pattern is a feasible way for diagnosis.

Idea: Build one tumor type classifier using gene expression data from all the tumor types.

Current differential analysis using RNA-Seq

Differential analysis is almost the most common practice in current RNA-Seq analysis.

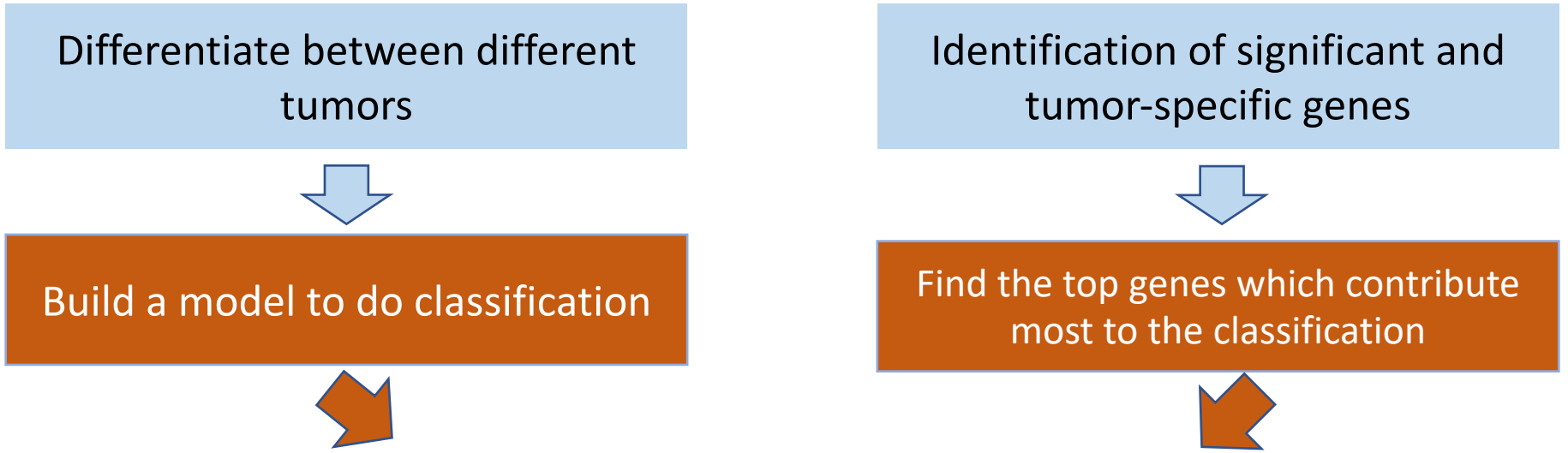


(1) Case/control samples correspond to the same tumor type.

(2) The candidate biomarkers found in this way may be shared by other tumor types, because no prior knowledge is applied. But right now we have this Pan-Cancer Atlas.

→ Discover tumor-specific genes.

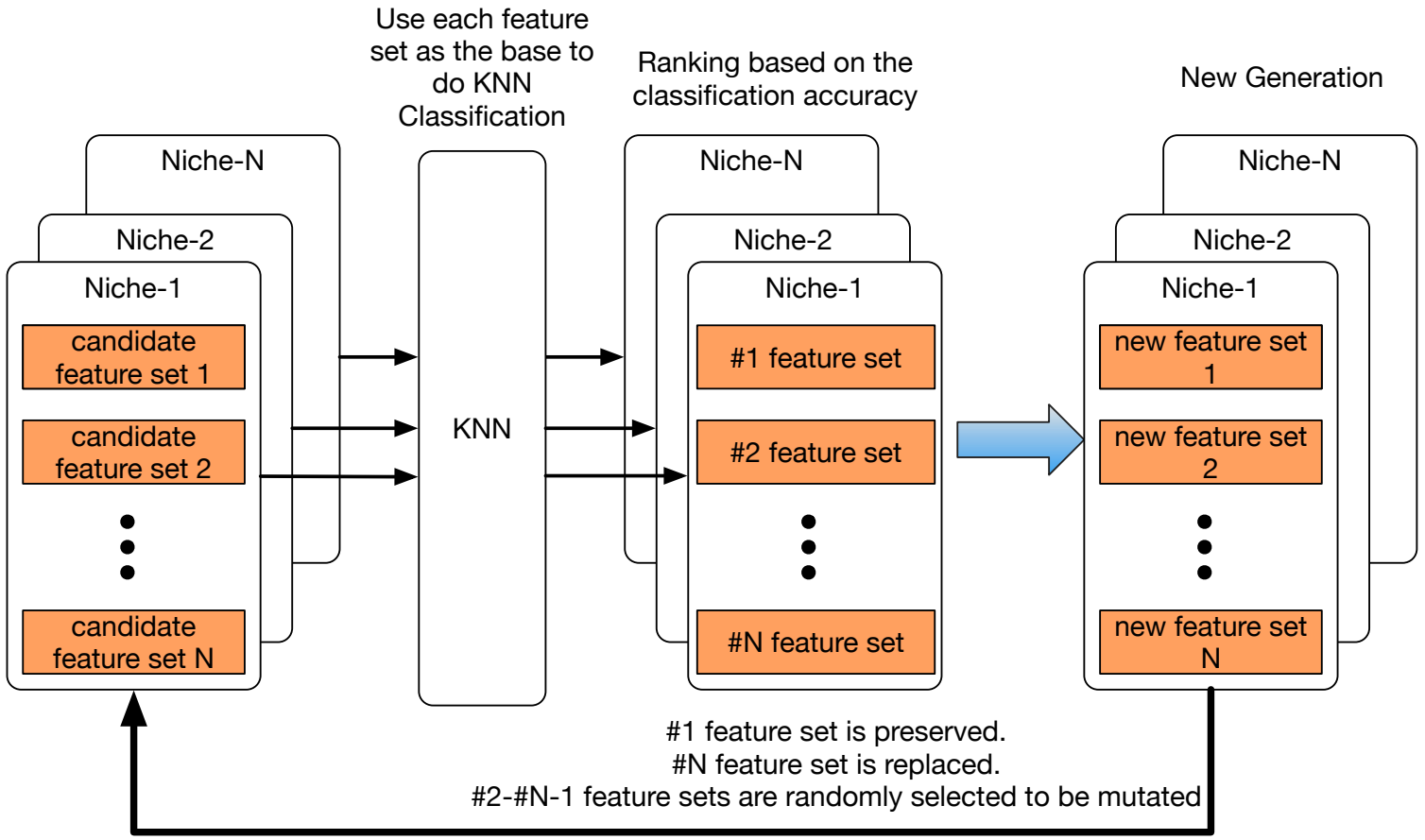
Research Problems



Reference method : GA/KNN[1]

Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M. and Li, L., 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, *18*(1), p.508.

GA/KNN used in the reference paper



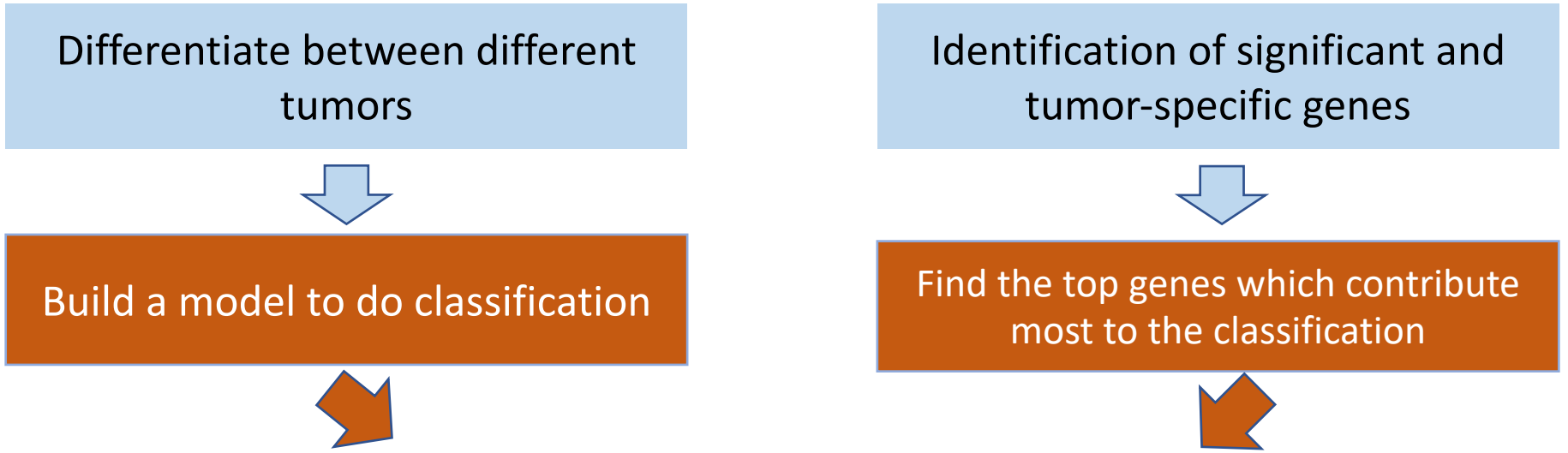
Idea: Iteratively build the feature set and search for the one with the best accuracy.

Pros:
Can obtain an optimal feature set through iterations.

Cons:
(1) The size of the feature set was fixed to 20.
(2) Fail to consider the dynamics of tumor-specific genes.

Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M. and Li, L., 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, *18*(1), p.508.

Research Problems

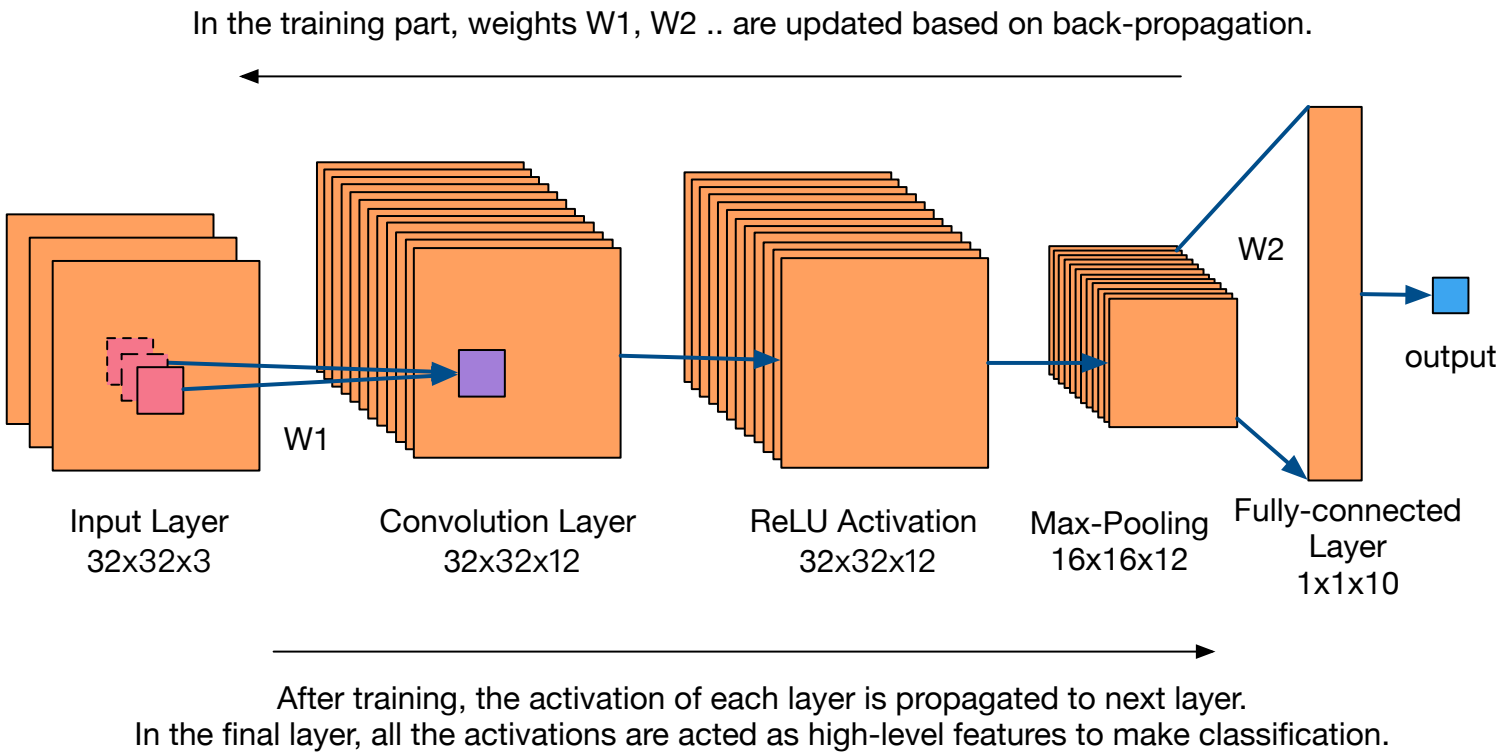


Our method : CNN + Guided Grad-Cam[2]

Guided Grad Cam: a technique to generate saliency map (heatmap of significance)

[2] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, *7*(8).

CNN

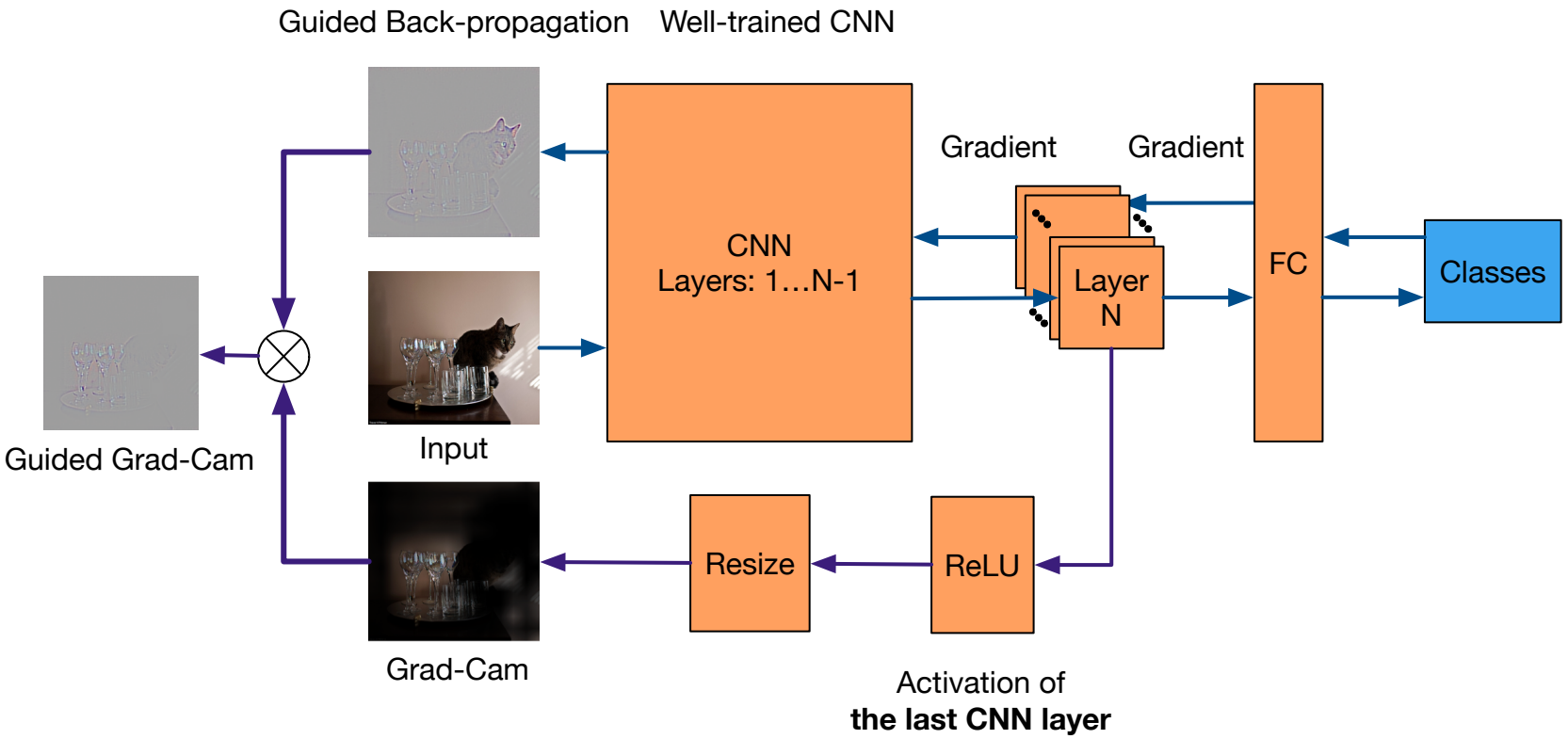


CNN: A feature extractor which transforms input into a high-level feature map.

The training part is actually tuning the weights of this feature extractor, so that the output is most close to the desired label.

These high-level features are significant to the classification.

Guided Grad-Cam



(1) **Higher layer contains higher level feature.**

(2) After the last convolutional layer, spatial information will be lost in FC layers.

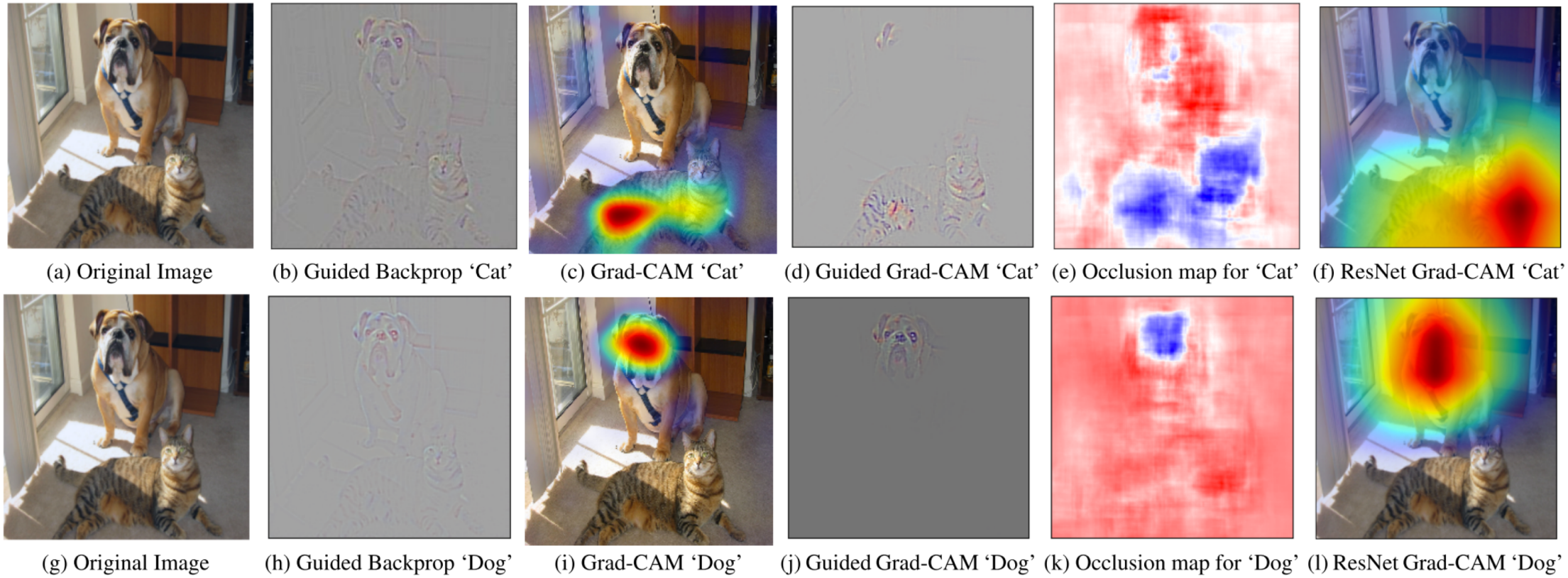
(3) Pixels with positive gradients have a relatively higher weight and also a higher significance to the next layer.

(4) **Grad-Cam: Localization map generated at the last convolutional layer.**

(5) **Guided Back-propagation: The positive gradients of the input back-propagated from the output.**

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, *7*(8).

Guided Grad-Cam

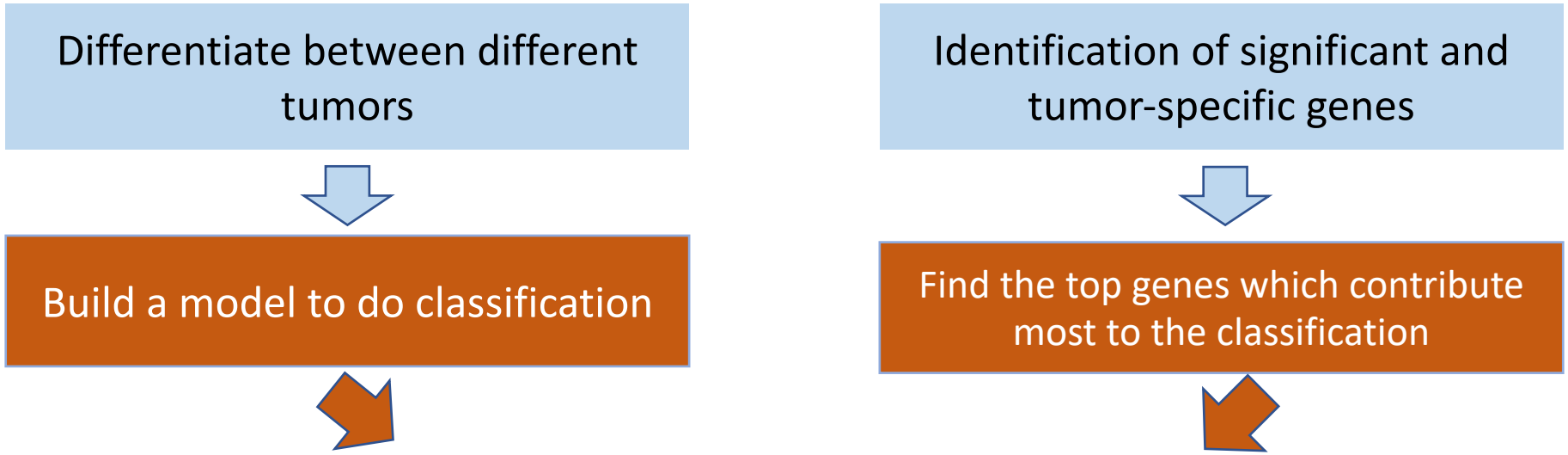


(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, *7*(8).

Overview of our method



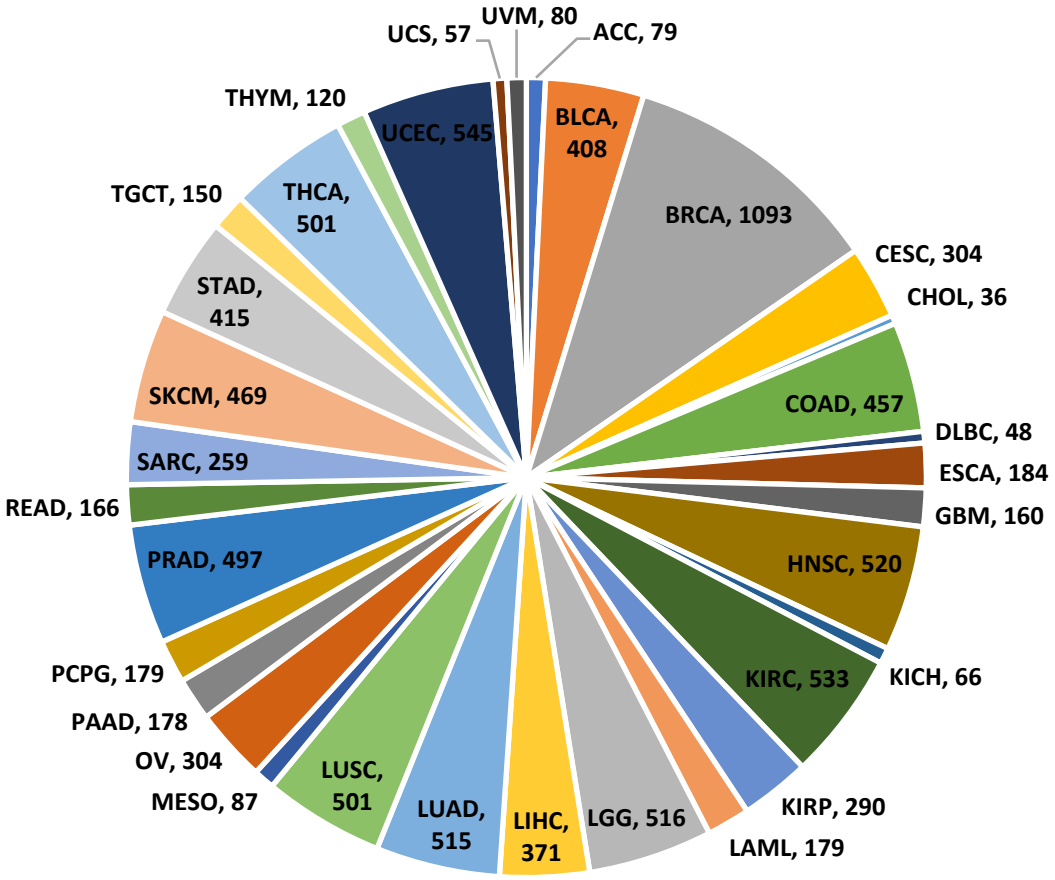
(1)Train a convolutional neural network for classification using gene expression data.

(2)By using the Guided Grad-Cam, we can also create a heatmap of significance in the input for each tumor type. The genes with higher intensities in heatmaps are considered as tumor-specific top genes. Because these genes contribute most to the classification of this tumor type.

In this way, two issues are solved at the same time by building one single model.

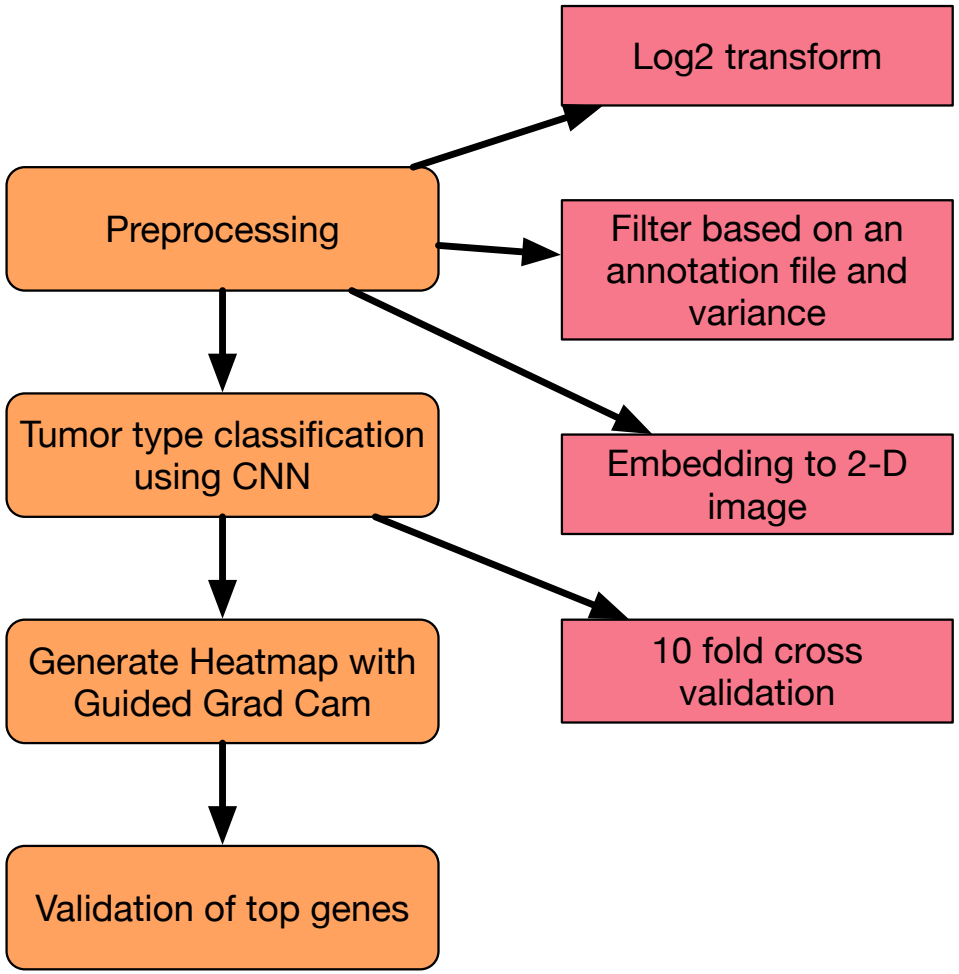
Data & method

- We use the analysis-ready standardized TCGA data from Broad GDAC Firehose. We use the **normalized-level3 mRNA-Seq expression data**.



(1) 10267 samples from 33 tumor types.
(2) Each sample contains normalized gene expressions for 20531 genes

Data & method



Preprocessing:

(1) All the values are first Log2-transformed (y = log2(x+1)), so that very large values are scaled to a smaller range to avoid bias.

(2) Set the values smaller than 1 to be zero, since they are very likely to be noise.

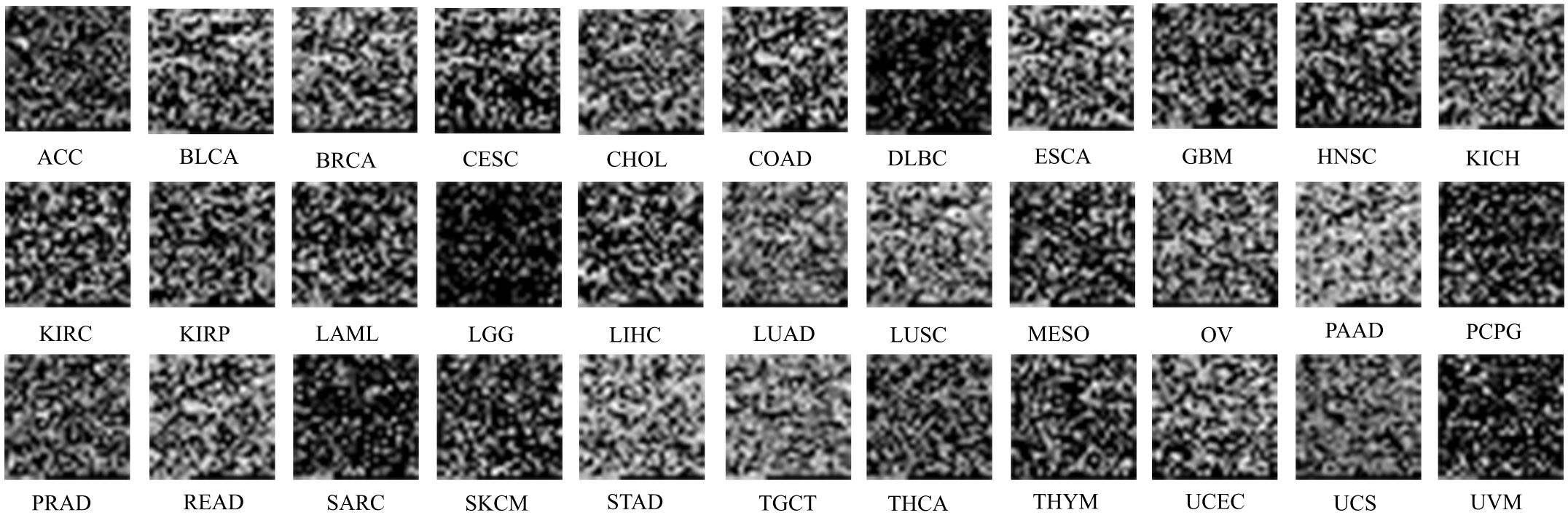
(3) Based on the homo sapiens gene annotation file, (updated 04/03/2018) to filter out the genes which are not in this annotation file. Downloaded from NCBI. ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

(4) Genes are filtered based on variance from # = 20531 to # = 10381, to remove noisy genes.

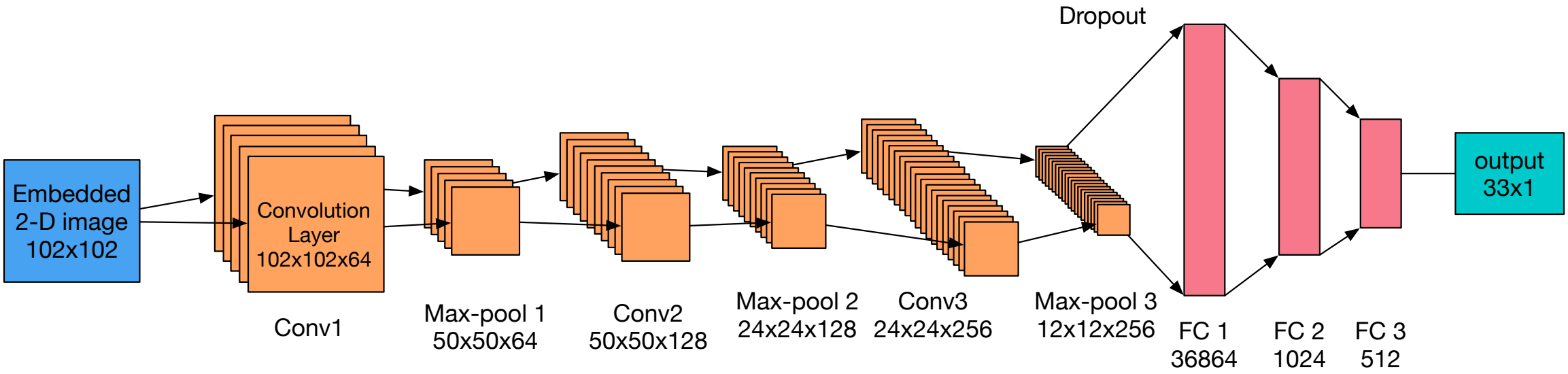
(5) Naively order the expression value based on the chromosome of corresponding gene. (chromosome numbers are from the annotation file.)

Preprocessing

- Embedding to 2-D image: we reshaped the **10381x1** array into a **102x102= 10404** gray image. Zero padding is added to the last several pixels. Also, the values are normalized to [0, 255]

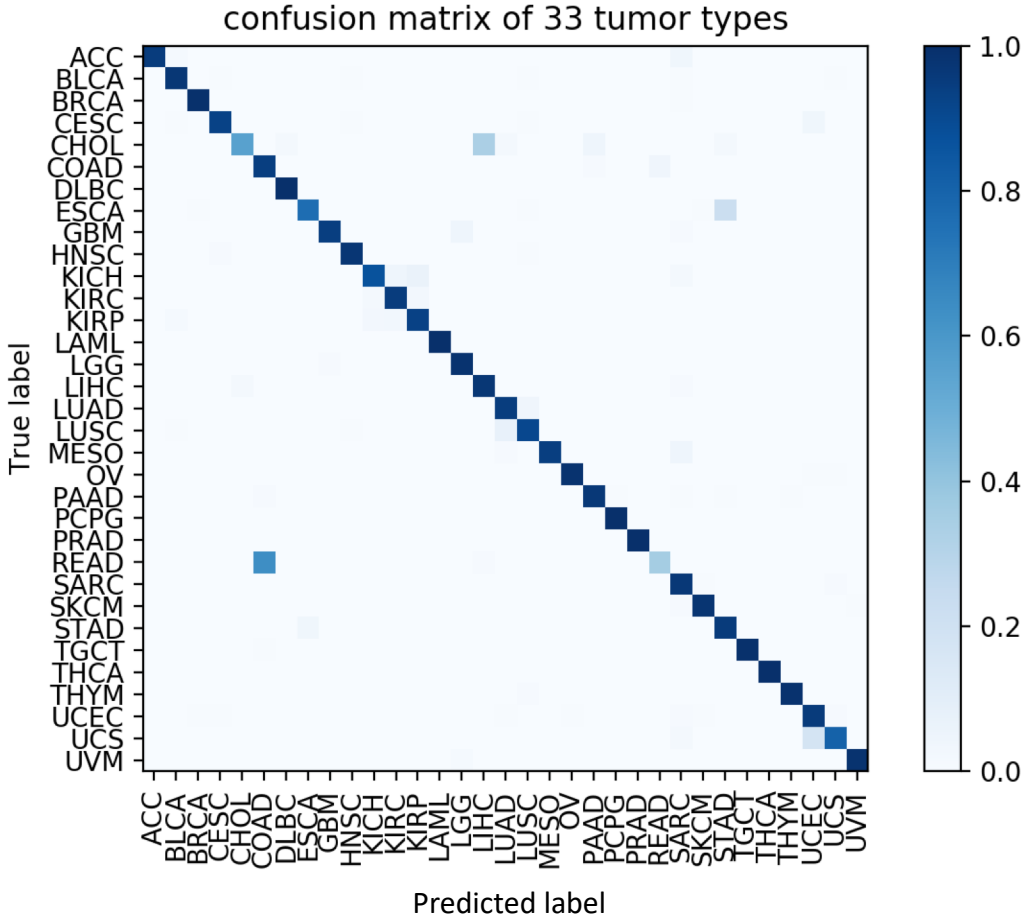


Training & testing

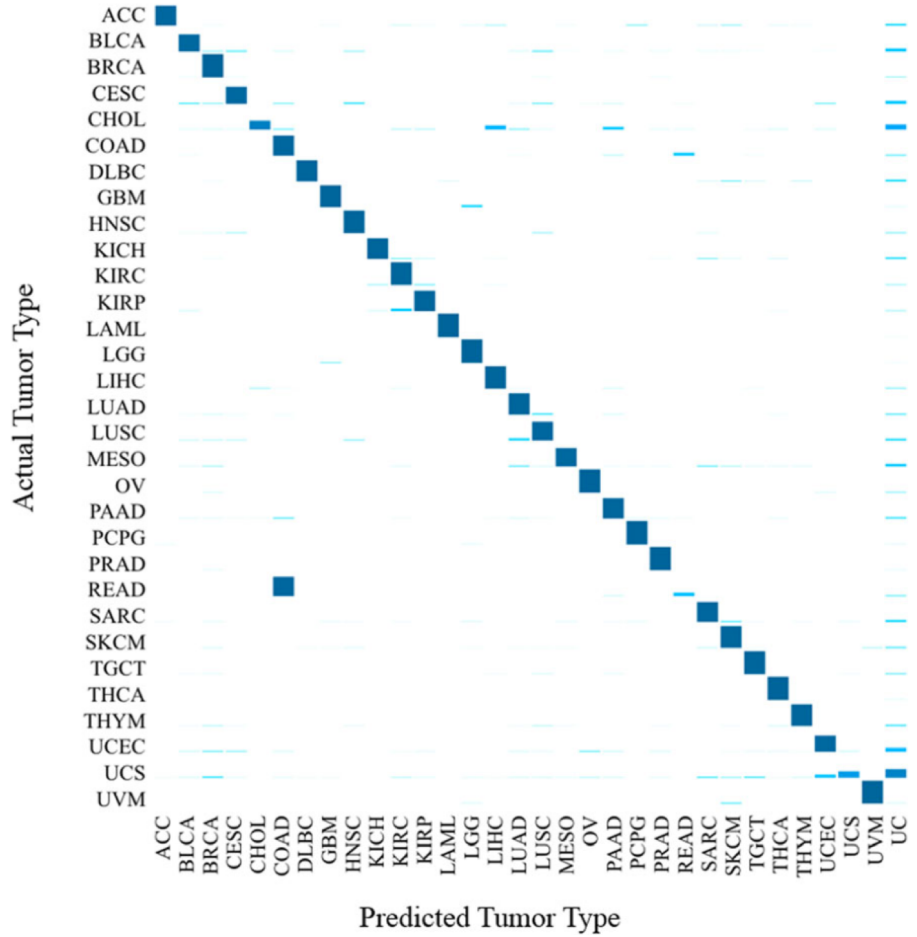


1. 3 convolutional layer neural network. Use cross entropy loss function and Adam optimizer during training.
2. Each batch contains 500 samples and go through the neural network 200 times (200 Epochs)
3. Early stopping was added during training to avoid overfitting.
4. Use Pytorch (A python package) to build this neural network.
5. This neural network is trained on the Newriver Cluster of Virginia Tech, using 2 Nvidia P100 GPU to run in parallel. Each Epoch took around 19s.

Accuracy: 95.59% (Reference : 90%)



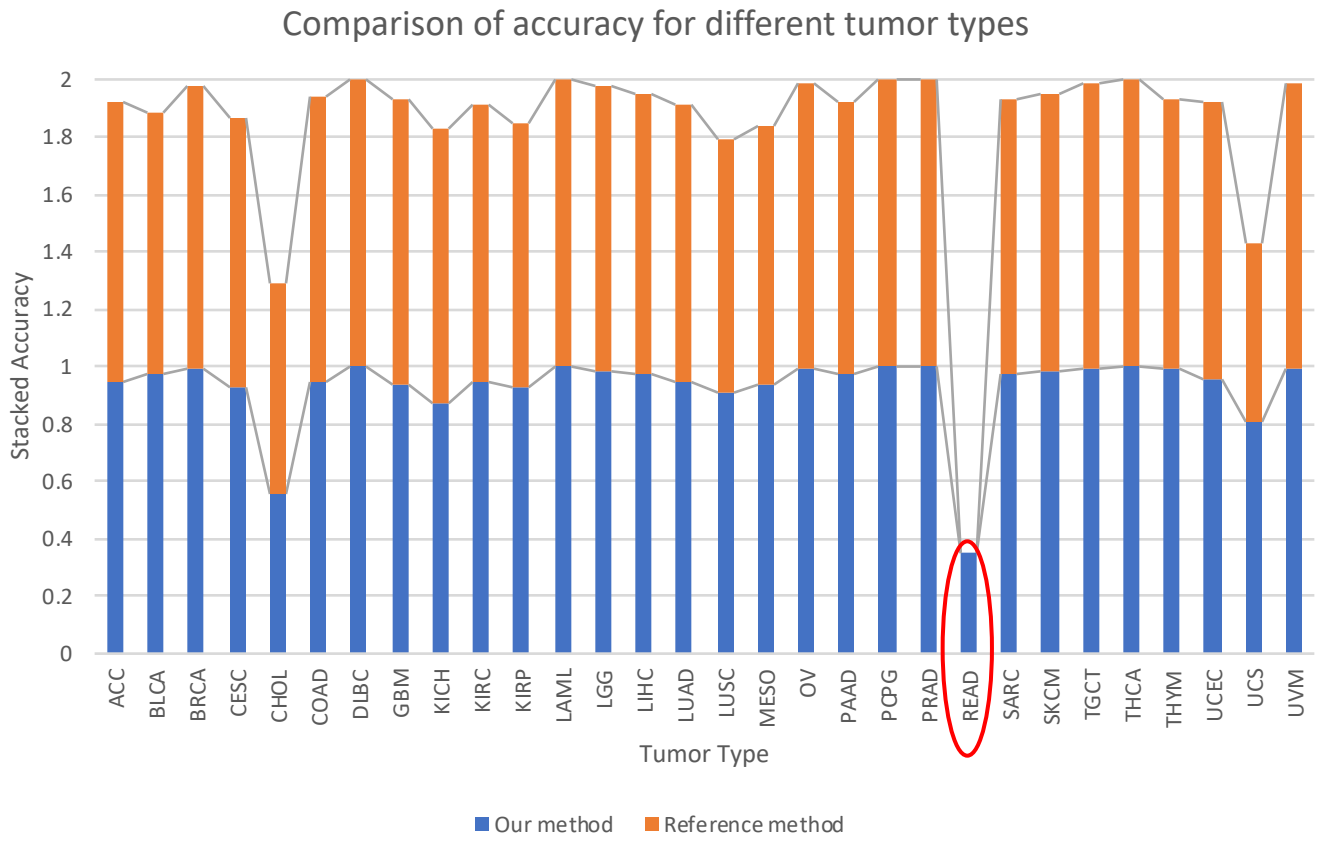
Our result



Result of the reference paper

Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M. and Li, L., 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, *18*(1), p.508.

Comparison with GA/KNN[1]



(1) An comparison of accuracy for each tumor type shows the superiority of our method.

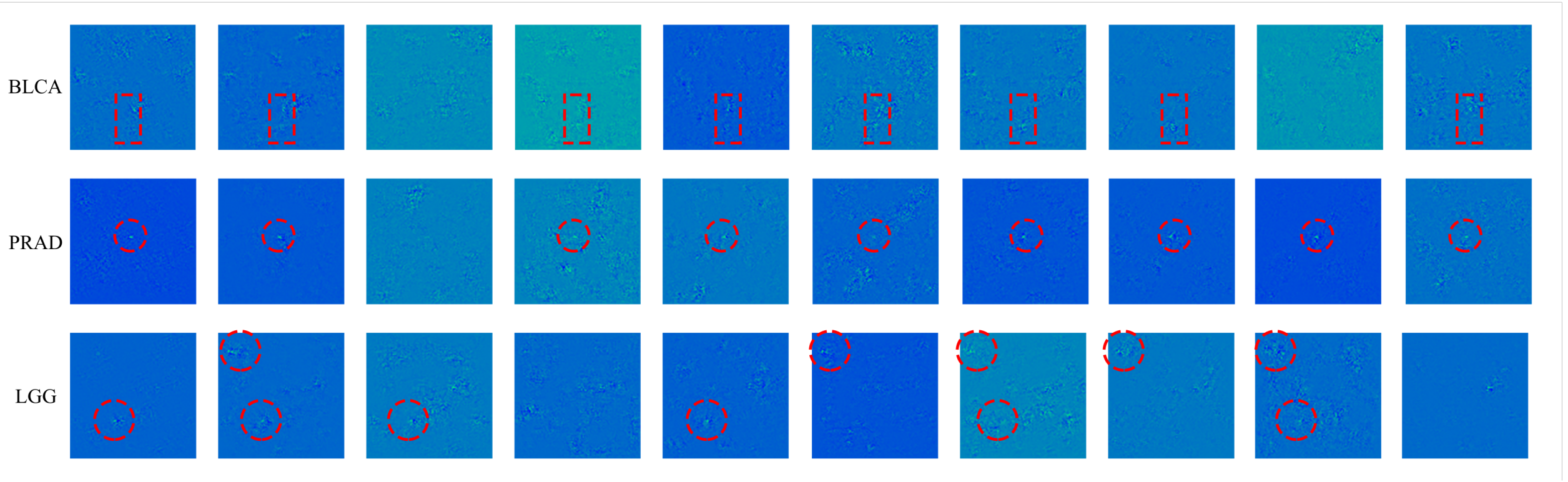
(2) Especially as to the tumor type READ, the accuracy in the reference was 0, while we got an accuracy of 35%.

(3) In addition, we did classification of all 33 tumor types, while the reference did classification over 31 tumor types.

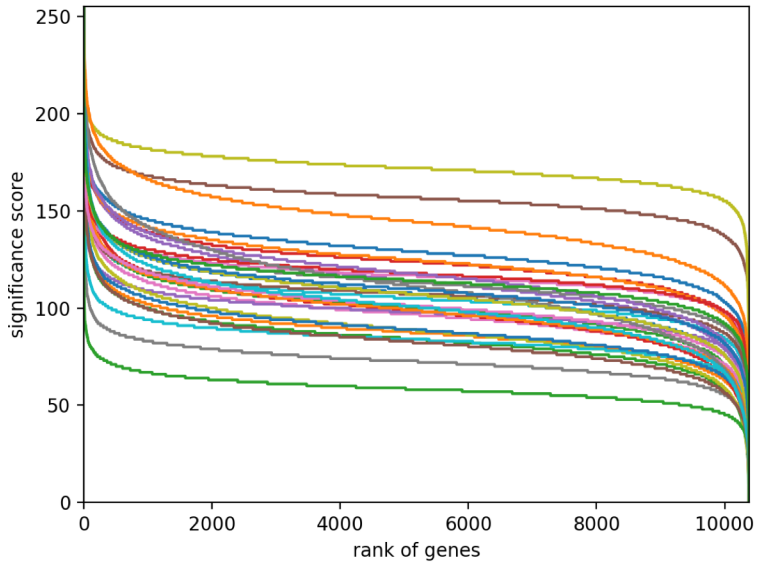
[1] Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M. and Li, L., 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, *18*(1), p.508.

Heatmaps for the significance of genes

- We used an open source guided grad cam to generate the heatmap.
- Steps: (1) With the trained network, as to each sample (image), forward pass each image, and obtain the activation map of the last convolutional layer. (2) Backpropagation to obtain the gradient of the last convolutional layer and each pixels of the input. (3) Multiply the activation and the gradients, and then normalized to [0,255]



Validation of genes



As to different classes, significance score from the heatmaps varies in a similar pattern with the rank of genes.

(1) Top 400 genes are selected based on the significance score in each classes.
(2) Pathway analysis is applied on these genes.
(3) Top genes of 16 tumor types are related to at least one pathway ($P < 0.05$), which corresponds to the tumor types.
(4) Concurrent pathways are found from top genes of 8 other tumor types. However, the related genes are different, which can be potential biomarkers.
(5) In the other 9 tumor types, no significant enriched pathways are found but their top genes are found to be related to the tumor types based on the GeneCards website.

Rank	COAD	GBM	LGG	LUSC	OV	UVM
1	LGALS4	GFAP	HSPB1P1	SFTPA2	MUC16	CD44
2	FCGBP	CBR1	HNRNPA1P33	KRT6A	KLK6	LGALS3BP
3	FTL	LOC613037	EEF1A1P9	SFTPA1	KLK7	SERPINF1
4	HOXC6	TIMP2	FTHL3	SFTPB	KLK8	GAPDHS
5	IGFBP2	IGFBP5	GFAP	KRT6B	KCNK15	MGST3

Validation of genes

Rank	COAD	GBM	LGG	LUSC	OV	UVM
1	LGALS4	GFAP	HSPB1P1	SFTPA2	MUC16	CD44
2	FCGBP	CBR1	HNRNPA1P33	KRT6A	KLK6	LGALS3BP
3	FTL	LOC613037	EEF1A1P9	SFTPA1	KLK7	SERPINF1
4	HOXC6	TIMP2	FTHL3	SFTPB	KLK8	GAPDHS
5	IGFBP2	IGFBP5	GFAP	KRT6B	KCNK15	MGST3

- As to COAD, the expression of its top1 gene LGALS4 (Galectin 4) is restricted to small intestine, colon, and rectum.
- As to GBM (Glioblastoma multiforme), a cancer in brain region, its top1 gene GFAP (Glial Fibrillary Acidic Protein) is used as a marker to distinguish astrocytes from other glial cells during development, which is also in the brain region.
- LGG (Brain Lower Grade Glioma) is also a tumor in the brain, its top1-4 genes are all pseudo-genes, while the top5 gene GFAP is the gene related to brain.
- As to LUSC (Lung squamous cell carcinoma), its top1 gene SFTPA2 has been implicated in many lung diseases
- As to OV (Ovarian cancer), its top1 gene MUC16 (CA125) was said to be the only reliable diagnostic marker for ovarian cancer [4].
- As to UVM (Uveal Melanoma), its top1 gene CD44 were tested to be strongly expressed in several cell lines of human uveal melanoma

[4] Mildred Felder, Arvinder Kapur, Jesus Gonzalez-Bosquet, Sachi Horibata, Joseph Heintz, Ralph Albrecht, Lucas Fass, Justanjyot Kaur, Kevin Hu, Hadi Shojaei, et al. 2014. MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. Molecular cancer 13, 1 (2014), 129.

Conclusion

- Using the gene expression data, we built a convolutional neural network to do classification for tumor types.
- And by applying the Guided Grad Cam technique on the trained CNN, we were able to assign each gene with a significance score.
- Genes with high significance scores in each tumor type are considered as top genes. Pathway analysis and survey of these top genes prove that they can be potential tumor-specific biomarkers.

Thanks!

Q & A