Deep Learning Based Tumor Type Classification Using Gene Expression Data

Boyu Lyu
Virginia Tech
Blacksburg, Virginia
boyu93@vt.edu

Anamul Haque
Virginia Tech
Blacksburg, Virginia
anamulmb@vt.edu

ABSTRACT

Differential analysis occupies the most significant portion of the standard practices of RNA-Seq analysis. However, the conventional method is matching the tumor samples to the normal samples, which are both from the same tumor type. The output using such method would fail in differentiating tumor types because it lacks the knowledge from other tumor types. Pan-Cancer Atlas provides us with abundant information on 33 prevalent tumor types which could be used as prior knowledge to generate tumor-specific biomarkers. In this paper, we embedded the high dimensional RNA-Seq data into 2-D images and used a convolutional neural network to make classification of the 33 tumor types. The final accuracy we got was 95.59%, higher than another paper applying GA/KNN method on the same dataset. Based on the idea of Guided Grad Cam, as to each class, we generated significance heat-map for all the genes. By doing functional analysis on the genes with high intensities in the heat-maps, we validated that these top genes are related to tumor-specific pathways, and some of them have already been used as biomarkers, which proved the effectiveness of our method. As far as we know, we are the first to apply convolutional neural network on Pan-Cancer Atlas for classification, and we are also the first to match the significance of classification with the importance of genes. Our experiment results show that our method has a good performance and could also apply in other genomics data.

KEYWORDS

Deep Learning, Tumor Type Classification, Pan-Cancer Atlas, Convolutional Neural Network

1 INTRODUCTION

The invention of Next Generation Sequencing methods has mostly boosted the analysis of human genomics due to the improvement in the efficiency and accuracy. To better understand the cause of various tumors, a large volume of tumor tissues has been sequenced and managed by The Cancer Genome Atlas (TCGA). With these samples, TCGA further analyzed over 11,000 tumors from 33 most prevalent forms of cancer, which fostered the accomplishment of Pan-Cancer Atlas. Where, as to each tumor sample, we could access its RNA-Seq expression data. These data are beneficial when trying to identify potential biomarkers for each tumor.

As to biomarkers, most analyses tried to find the differentially expressed genes. However, they didnt consider expressions of other tumors. Also, during the study, models are constructed to mimic the expression data. However, such models are very data specific, which would fail in dealing with data from other samples or other tumor types. So, it is highly needed to build a method which can include knowledge from multiple tumor types into the analysis.

On the other hand, tumor type classification can contribute to a faster and more accurate diagnosis. However, research on tumor type classification is currently rare, which is partially due to the difficulty in dealing with high dimensionality of the genomic dataset. In Pan-Cancer Atlas, the normalized mRNA-seq gene expression data contains information from more than 20k genes. Within that many genes, a lot of genes are actually of no specific effect on the tumor. These genes are weak features. Using generic machine learning methods such as KNN wont work due to the curse of dimension. Even though classification of tumor type is still in its beginning, deep learning has been vastly used in image classification/ recognition, some famous architectures like Resnet [9] and inception[17] have excellent performance. In the Imagenet 2017 challenge, the winner obtained a top-5 error rate of 2.251%[10] using a convolutional neural network. Besides, to understand how deep neural network works, in computer vision field, various visualization methods have been invented such as deep Taylor decomposition, layer-wise decomposition, and Grad-Cam[16]. They generate heat-maps in the input image to indicate the contribution of each pixel to the classification. Since the gene expression data contains more than 10k samples in the atlas, it is promising to have good accuracy with the deep neural network. At the same time, assuming that the importance of genes equals to the significance of genes in classification, we could borrow the interpretation methods of the deep neural network to discover top genes for each tumor. In this way, each gene will be assigned a confidence score, and genes with high scores are considered to be the top genes or potential biomarkers since their existence affect the classification most.

Based on the deep neural network and the visualization methods, we first filter out the genes with small variance across all the samples. Then, we embedded the high-dimension expression data (10381x1) into a 2-D image (102x102) to fit for the convolutional layers. Next, we constructed a three-layer convolution neural network and used 10-fold cross-validation to test the performance. With the trained neural network, we followed the idea of Guided

ACM-BCB '18, August 2018, Washington, D.C., USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Grad-Cam[16] and generated heat-maps for all the classes showing prominent pixels (genes). By functional analysis, we validated that the top genes selected in this way are biologically meaningful for corresponding tumors, which proves the effectiveness of our work.

The remaining of this paper is organized as follows. In section 2, we reviewed the related work of tumor type classification and visualization of the deep neural network. In section 3, we described the data we used, and the process of we applied for tumor type classification. Also, the steps to generate heat map for each sample are also explained. Finally, in section 4, The accuracy of the tumor type classification is discussed and compared with the result of GA/KNN method in paper[11]. Also, top genes are picked from the heat-maps and validated by functional analysis.

2 RELATED WORK

Binary classification (identification) of tumors has been found in some papers, but only paper [11] researched multiple tumor type classification using gene expression data. The authors applied GA/KNN method to iteratively generate the subset of the genes (features) and then use KNN method to test the accuracy. After hundreds of iterations, they obtained the feature set corresponding to the best accuracy. In this way, this method achieved an accuracy of 90% across 31 tumor types, and also generated a set of top genes for all the tumor type. GA/KNN method is a straightforward method which could obtain an optimal feature set in the end, but it requires running a lot of iterations and also fixes the size of feature set at first. However, using one single feature set to make classification overlooked the heterogeneity among different tumor types. Also, top genes for various tumor types could be varied a lot.

Deep learning method was also used to identify top genes and make cancer identification of one single type. In paper [5], the author used a stacked auto-encoder first to extract high-level features from the expression values and then input these features into a single layer ANN network to decide whether the sample is a tumor or not. The accuracy using such method reached 94%. However, as to multiple classes, this method will lead to a very complicated structure, since this is not an end-to-end method. To identify the top genes of BRCA, weight matrices of each layer in the auto-encoder are multiplied to obtain the estimated weights for each gene at the input layer. By fitting with the normal distribution, they selected the top genes. This idea is similar to our view of using visualization methods. The difference is that they matched the importance of genes to the high-level features, whereas we matched the importance of genes to their contribution to the classification.

Deep Taylor decomposition and layer-wise relevance propagation (LRP) [1, 2] are designed to interpret the deep neural network by backpropagation. Where one conservative way to perform LRP is to redistribute the input of each neuron back to all of its predecessors equally. And layer by layer, when the input layer is reached, the decomposition is done. However, LRP methods require a relevance conservation property between layers, which puts some constraints on the neural network. While another technique, guided Grad-Cam[16], a combination of Guided back-propagation and Grad-Cam, can be applied on any neural network without any modification to the original architecture. In this paper, we are going to use

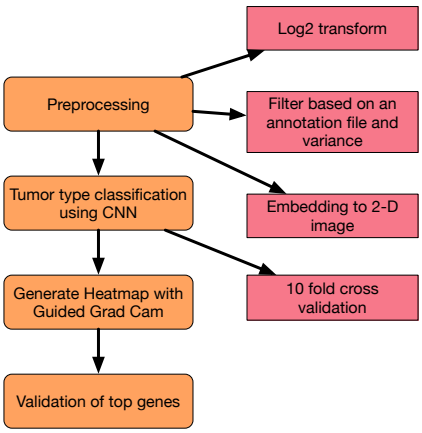

Figure 1: The workflow of our method

this method to trace the significant pixels (genes) in the input, so that we could extract substantial genes. The localization map of Grad Cam is generated by first calculating the activation map in the forward propagation and the gradient map in the backward propagation. Since the deeper the convolutional layer, the higher level feature it contains. And after sending into the fully connected layer, all the spatial information will be lost. So, Grad-Cam constructs the localization map of the last convolutional layer and resizes it to the input size. However, as stated by the [5], the heat map shows the importance of each class, but the resolution is low, due to the resizing process. To solve this problem, a pixel-space gradient visualization method, guided back propagation, was used to refine the heat map. The output of guided backpropagation is the gradient of each pixel in the input image, which can be obtained through the backpropagation. Then the final output of this Guided Gram-Cam can be obtained by multiplying the results of guided backpropagation and Grad-Cam.

3 DATA AND METHOD

3.1 Classification

The workflow of our method is shown in figure1. (1) Preprocessing of the input data. (2) Make tumor type classification using a convolutional neural network. (3) Generate the heat map for each class and pick the genes corresponding to top intensities in the heat-maps. (4) Validate the pathways of selected genes.

3.2 Data

We used the normalized-level3 RNA-Seq gene expression data of 33 tumor types in Pan-Cancer Atlas. A detailed list of all 33 tumor types and corresponding number of samples is shown in table 3. The data contains 10267 tumor samples with respect to 20531 genes.

3.3 Preprocessing

The data contains the normalized read count for each gene, but the range of the values is enormous, and also some values are smaller

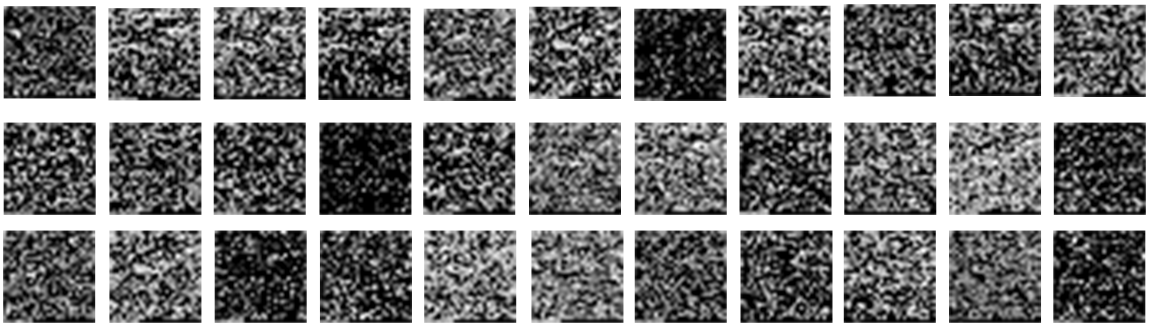

Figure 2: An example of embedded 2-D images. From top left to the bottom right are the images generated from class 1 to class 33.

than 1. So we first apply a transform using $y = \log_2(x + 1)$, and set all the values lower than 1 to be 0 since these values are very likely to be noise. Besides, we matched the genes with an annotation file (updated on 04/03/2018) downloaded from NCBI, around 1000 genes were not found in the annotation file so that we removed these genes and corresponding expression level. Then we used a variance threshold of **1.19** to filter out the genes of which expression levels remain almost unchanged, and the number of genes is further reduced to 10381.

To input the data to the convolutional neural network, we embedded the data into 2-D images. First genes are ordered based on the chromosome number because adjacent genes are more likely to interact with each other. Then the data is reshaped from a 10381×1 array into a 102×102 image by adding some zeros at the last line of the image. Then all the images are normalized to make sure that the range is [0,255]. The generated images of different classes are shown in figure 2.

3.4 Classification

We used a convolutional neural network consisting of three convolutional layers, and three fully connected layer, which is shown in figure 5. The first convolutional layer 'conv1' contains 64 different filters, while the second and the third convolutional layers contain 128 and 256 filters respectively. Max-pooling layer and batch-normalization layer are followed immediately after each convolutional layer. A drop-out layer is added before entering into fully connected layer, the drop-out rate is 25%. The size of the three fully connected layers are 36864, 1024, 512 separately. We chose Cross Entropy as loss function and Adam optimizer to update the weights. We used 10-fold cross validation to train the convolutional neural network and to test the performance.

3.5 Heat-map

Based on the idea of Guided Grad-Cam, we let the training data set go through the trained neural network. We recorded the activation maps in the forward pass and the gradient maps in the backpropagation, then we could obtain the heat map for each sample using Guided Grad-Cam. And by averaging all the heat-maps from the same class and after normalization, we could obtain the class specific heat-map. As to each pixel in the heat-map, a higher intensity represents a higher contribution to the final classification, which indicates a higher importance of its corresponding gene.

In fact, this heat-map generating process could be implanted into the training of the convolutional neural network, because it will not affect the training and testing process. But the computation is less using the above two-step method, because Grad-Cam only requires all the samples to run through the neural network once.

3.6 Validation

Top genes are selected based on the intensity rankings in the heatmaps. We applied functional analysis on these top genes to further prove that the genes are tumor specific and are potential biomarkers. In the first stage, we chose top 400 genes of each tumor type to do pathway analysis, trying to find out if significantly enriched pathways are related to the corresponding tumor. In the second stage, we studied the top 5 genes for the tumor types presenting few results in the first stage, to find their relations to the tumor.

4 EXPERIMENT RESULT

4.1 Classification

Using 10-fold cross validation, we calculated the average accuracy, average accuracy of each class, average precision score P and average recall score R, and average F1 score $F1$. Where,

$$P = \frac{TP}{TP + FP}, \; R = \frac{TP}{TP + FN}, \; F1 = \frac{2PR}{P + R} \tag{1}$$

The scores are summarized into table 1

Table 1: Performance of our method

method	accuracy	precision	recall	f1-score
CNN	95.59%	95.54%	95.59%	95.43%

In addition, we have also generated a confusion matrix as shown in figure3 We can see most classes are classified correctly, but there are two pairs of misclassification. (1) READ samples are mostly

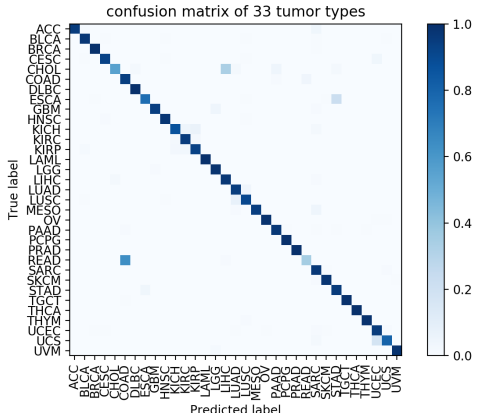

Figure 3: The confusion matrix generated using our method

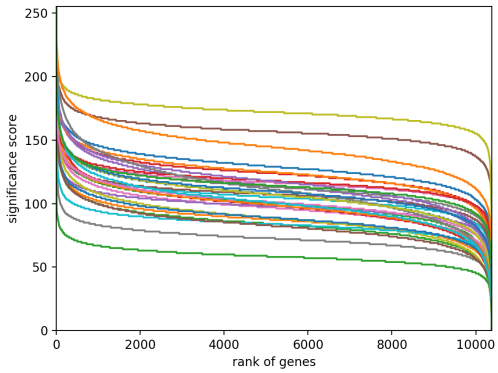

Figure 4: The changes of intensities in heat-maps for each class. It can be seen that different classes shared the same pattern of change in intensities.

misclassified into COAD which might be due to the close spatial relationship between the two tumor types. (2) Some CHOL samples are misclassified into LIHC due to the small number of CHOL samples. But using CNN do show an improvement in dealing with READ samples compared to the reference[11]. A comparison of accuracy as to each class is shown in table 3. From the table we can see that our accuracies in dealing with classes READ and UCS have been improved a lot compared with the reference paper.

4.2 Generated Heat-map

The heat-maps generated for each class showed a similarity across ten folds, and displayed a distinct pattern when comparing in-between classes. Some examples are shown in figure 6. In the example, each row represents heat-maps of one tumor type, and columns represent heat-maps from different folds (from left to right are fold 1 to fold 10). Circled regions show the similar pattern in all the folds. Even though there are some difference among different folds, there do exist a clear pattern.

4.3 Validation of Top Genes

In the heat-maps, we found that the intensities of top 100 genes decreased sharply, while the intensities decreased very smoothly starting from around 400th genes. When the intensities are very small, the decrease rate became high again. Such change can be viewed in fig 4. Assuming that the larger intensity implies more significance, since the slopes of the intensity change in the first 400 genes are larger than the following several thousands genes, we chose the top 400 genes as query genes. Besides, comparing to the total number of genes (10381), our choice of 400 is still a very small number, which is consistent with the fact that the number of biomarkers should be small. The KEGG pathway analysis results for top 400 genes of each tumor are obtained using the David website[1]. Next, we did a literature review on the results trying to find out the relations between these pathways and tumor types. The related pathways with P value smaller than 10^{-3} are shown in table4. Where the bold ones are related to corresponding tumor types. The top 400 genes of 24 tumor types were found to have significant enrichments of pathways. In which, 16 tumor types have at least one related pathway given the top 400 genes. The related genes in these pathways can then be viewed as tumor specific biomarkers.

While as to the other 8 tumor types, even though no direct related pathways are found, two significant enriched concurrent pathways are discovered, which are hsa04512: ECM-receptor interaction and hsa04510: focal adhesion. However, the related genes are not the same as to different tumor types. For example, in BRCA, ECM-receptor interaction pathway related genes are SDC1, COL4A1, COL3A1, COL6A3, COL1A2, ITGA11, COL6A2, COL6A1, COL1A1, COL5A2 and FN1. Whereas in ESCA, related genes are VWF, ITGA6, LAMC3, ITGAV, TNC, ITGB6, ITGB4, ITGA3, COL1A1, COL5A2 and COL4A6. Similarly, in terms of focal adhesion pathway, related top genes in SKCM are CAV1, TNC, COL3A1, PIK3CD, ITGA3, ITGB3, COL5A3, CCND1, LAMB2, FYN, COMP, COL6A3, COL6A2, COL6A1, COL1A1, LAMB1, PARVB, SPP1, FN1 and SHC4. Whereas related genes in THCA are COL4A4, COL4A3, CAV2, CAV1, PGF, BCAR1, MET, IGF1, ITGA3, LAMA5, COL1A2, LAMC2, COL1A1 and FN1. It can be seen that these sets are not quite the same in different tumor types, which shows that they can be potential biomarkers.

In the other 9 tumor types (CHOL, COAD, READ, GBM, KICH, LGG, LUSC, OV and UVM), no significant enriched pathways were found. The samples for CHOL and KICH are limited, so we omitted these two tumor types. Besides, the classification accuracy of class READ is very low, so we also omitted it. With respect to the 6 remaining tumor types, we reduced the query size from top 400 to top 5 . The query genes are shown in table 2. We looked into their information on the GeneCards website [2]. As to COAD, its top1 gene LGALS4 (Galectin 4) is a Protein Coding gene. The expression of this gene is restricted to small intestine, colon, and rectum, and it is under-expressed in colorectal cancer. As to GBM (Glioblastoma multiforme), a cancer in brain region, its top1 gene GFAP (Glial Fibrillary Acidic Protein) is a Protein Coding gene. This gene encodes

[1]David:https://david.ncifcrf.gov/home.jsp

[2]GeneCards: www.genecards.org

Table 2: Top5 genes for the 6 remaining tumor types

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

one of the major intermediate filament proteins of mature astrocytes. It 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 [13]. As to OV (Ovarian cancer), in paper[7], its top1 gene MUC16 (CA125) was said to be the **only** reliable diagnostic marker for ovarian cancer. As to UVM (Uveal Melanoma), its top1 gene CD44 were tested to be strongly expressed in several cell lines of human uveal melanoma[4]. All the above results have shown the top genes have very close relations to the corresponding tumor types, which could be viewed as potential biomarkers.

5 DISCUSSION

Genomic tests are becoming popular nowadays. With saliva or blood, these tests can tell out the estimated possibilities of different tumors for each individual. Such tests mostly rely on biomarkers of various tumors. However, gene expression biomarkers generated by differential analysis are not guaranteed to be tumor specific, since several tumors might share the same biomarkers. To avoid this problem, we designed a method using the knowledge from multiple tumor types and found the genes which can be used to differentiate between different tumor types. Validation results prove that these top genes are possible to be biomarker due to their relations to the corresponding tumor type. Further examinations can be made biologically.

In this paper, we used a convolutional neural network to make classification on the genomics data. Research on the computer vision is developing very fast. A lot of methods were designed to solve problems using the deep neural network. However, in the genomics community, the research based on deep learning is still in its beginning. One issue is that genomics data usually are high dimensional, while most deep learning architectures are for 2-D images. While, we have shown that, by just naively placing the genomics data (genes) onto each pixel of the 2-D image based on the order of chromosome number, the performance was excellent except for several tumor types. So that, it is possible that many more deep learning methods could be applied to the genomics.

Looking into the misclassified samples, we think one possible issue is the imbalanced dataset. Some tumor type has over 1000 samples while some only has 30 samples. Imbalanced dataset might cause a big problem in deep learning, so for a better result, this problem can be mitigated using oversampling methods such as SMOTE.

6 CONCLUSION

In this paper, we designed a new method to discover potential biomarkers for each tumor type. We matched the importance of genes to the contribution to the correct classification. Based on the Pan-Cancer Atlas, we used a convolutional neural network to make the classification and used a visualization method of neural network to discover top genes from the input. The accuracy of our method shows improvement compared with previous work on tumor type classification. And we examined the top genes of each tumor type, found their relations to the corresponding tumor types, which proved that the top genes are potential biomarkers.

REFERENCES

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[2] Sebastian Bach, HHI FRAUNHOFER, Alexander Binder, EDU SG, and Wojciech Samek. [n. d.]. Deep Taylor Decomposition of Neural Networks. ([n. d.]).

[3] MR Bishop, RM Dean, SM Steinberg, J Odom, SZ Pavletic, C Chow, S Pittaluga, Claude Sportes, NM Hardy, J Gea-Banacloche, et al. 2008. Clinical evidence of a graft-versus-lymphoma effect against relapsed diffuse large B-cell lymphoma after allogeneic hematopoietic stem-cell transplantation. *Annals of oncology* 19, 11 (2008), 1935–1940.

[4] WM Creyghton, EH Danen, GP Luyten, MJ Jager, et al. 1995. Cytokine-mediated modulation of integrin, ICAM-1 and CD44 expression on human uveal melanoma cells in vitro. *Melanoma research* 5, 4 (1995), 235–242.

[5] Padideh Danaee, Reza Ghaeini, and David A Hendrix. 2017. A deep learning approach for cancer detection and relevant gene identification. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*. World Scientific, 219–229.

[6] Panagiota Economopoulou, Ioannis Kotsantis, and Amanda Psyrri. 2016. The promise of immunotherapy in head and neck squamous cell carcinoma: combinatorial immunotherapy approaches. *ESMO open* 1, 6 (2016), e000122.

[7] 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.

[8] Todd M Gibson, Eric A Engels, Christina A Clarke, Charles F Lynch, Dennis D Weisenburger, and Lindsay M Morton. 2014. Risk of diffuse large B-cell lymphoma after solid organ transplantation in the United States. *American journal of hematology* 89, 7 (2014), 714–720.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507* (2017).

[11] Yuanyuan Li, Kai Kang, Juno M Krahn, Nicole Croutwater, Kevin Lee, David M Umbach, and Leping Li. 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics* 18, 1 (2017), 508.

[12] Rahil Mashhadi, Gholamreza Pourmand, Farid Kosari, Abdolrasoul Mehrsai, Sepehr Salem, Mohammad Reza Pourmand, Sudabeh Alatab, Mehdi Khonsari, Fariba Heydari, Laleh Beladi, et al. 2014. Role of steroid hormone receptors in formation and progression of bladder carcinoma: a case-control study. *Urology journal* 11, 6 (2014), 1968.

[13] Li Peng, Xiu Wu Bian, Chuan Xu, Guang Ming Wang, Qing You Xia, Qing Xiong, et al. 2015. Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Scientific reports* 5 (2015), 13413.

[14] Mark Schiffman, Philip E Castle, Jose Jeronimo, Ana C Rodriguez, and Sholom Wacholder. 2007. Human papillomavirus and cervical cancer. *The Lancet* 370, 9590 (2007), 890–907.

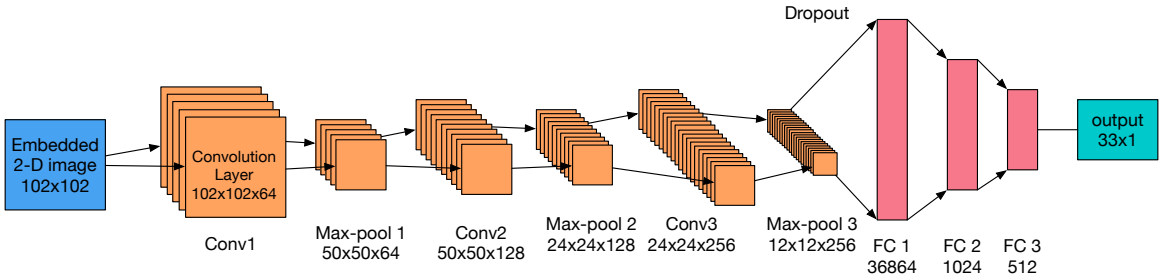

Figure 5: The architecture of our convolutional neural network.

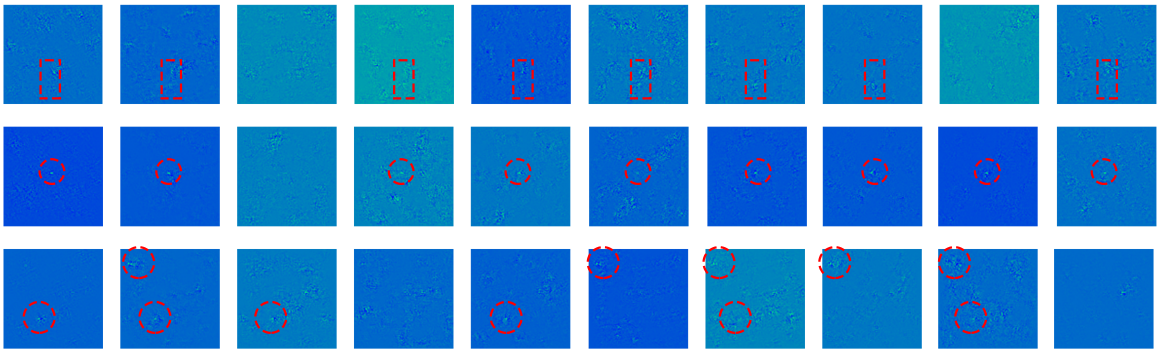

Figure 6: Some heat map examples. Each column represents the result from one fold. The first row are the heat-map of tumor type BLCA, the second row belongs to PRAD and the third row belongs to LGG.

[15] Yoshitaka Sekido. 2013. Molecular pathogenesis of malignant mesothelioma. *Carcinogenesis* 34, 7 (2013), 1413–1419.

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3* 7, 8 (2016).

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[18] Philip E Tarr, Michael C Sneller, Laura J Mechanic, Athena Economides, Christopher M Eger, Warren Strober, Charlotte Cunningham-rundles, and Daniel R Lucey. 2001. Infections in patients with immunodeficiency with thymoma (Good syndrome): report of 5 cases and review of the literature. *Medicine* 80, 2 (2001), 123–133.

[19] Chin-Hsiao Tseng and Farn-Hsuan Tseng. 2014. Diabetes and gastric cancer: the potential links. *World journal of gastroenterology: WJG* 20, 7 (2014), 1701.

[20] Gisele Moledo de Vasconcelos, Fernanda Azevedo-Silva, Luiz Claudio dos Santos Thuler, Eugênia Terra Granado Pina, Celeste SF Souza, Katia Calabrese, and Maria S Pombo-de Oliveira. 2014. The concurrent occurrence of Leishmania chagasi infection and childhood acute leukemia in Brazil. *Revista brasileira de hematologia e hemoterapia* 36, 5 (2014), 356–362.

Table 3: Tumor types and number of RNA-Seq samples

Tumor Type	Cohort	Number of samples	Accuracy of our method	Accuracy of reference
Adrenocortical carcinoma	ACC	79	0.95	0.97
Bladder urothelial carcinoma	BLCA	408	0.97	0.91
Breast invasive carcinoma	BRCA	1093	0.99	0.99
Cervical and endocervical cancers	CESC	304	0.93	0.94
Cholangiocarcinoma	CHOL	36	0.56	0.73
Colon adenocarcinoma	COAD	457	0.95	0.99
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	48	1.0	1.0
Esophageal carcinom	ESCA	184	0.77	None
Glioblastoma multiforme	GBM	160	0.94	0.99
Head and Neck squamous cell carcinoma	HNSC	520	0.98	None
Kidney Chromophobe	KICH	66	0.87	0.96
Kidney renal clear cell carcinoma	KIRC	533	0.95	0.96
Kidney renal papillary cell carcinoma	KIRP	290	0.93	0.92
Acute Myeloid Leukemia	LAML	179	1.0	1.0
Brain Lower Grade Glioma	LGG	516	0.98	1.0
Liver hepatocellular carcinoma	LIHC	371	0.97	0.98
Lung adenocarcinoma	LUAD	515	0.95	0.96
Lung squamous cell carcinoma	LUSC	501	0.91	0.88
Mesothelioma	MESO	87	0.94	0.90
Ovarian serous cystadenocarcinoma	OV	304	0.99	1.0
Pancreatic adenocarcinoma	PAAD	178	0.97	0.95
Pheochromocytoma and Paraganglioma	PCPG	179	1.0	1.0
Prostate adenocarcinoma	PRAD	497	1.0	1.0
Rectum adenocarcinoma	READ	166	**0.35**	**0.0**
Sarcoma	SARC	259	0.97	0.96
Skin Cutaneous Melanoma	SKCM	469	0.98	0.97
Stomach adenocarcinoma	STAD	415	0.96	None
Testicular Germ Cell Tumors	TGCT	150	0.99	1.0
Thyroid carcinoma	THCA	501	1.0	1.0
Thymoma	THYM	120	0.99	0.94
Uterine Corpus Endometrial Carcinoma	UCEC	545	0.96	0.96
Uterine Carcinosarcoma	UCS	57	**0.81**	**0.62**
Uveal Melanoma	UVM	80	0.99	1.0
Total		10267		

Table 4: Pathway analysis results on top 400 genes of each tumor type ($P < 10^{-3}$)

Tumor name	Related Pathway		P value
	ID	Name	
ACC	**hsa04925**	**Aldosterone synthesis and secretion**	3.13E-06
	hsa04913	Ovarian steroidogenesis	1.83E-05
	hsa01100	Metabolic pathways	4.69E-04
BLCA	**hsa00140**	**Steroid hormone biosynthesis[12]**	5.63E-07
	hsa04510	Focal adhesion	3.69E-06
	hsa04512	ECM-receptor interaction	3.71E-06
	hsa04974	Protein digestion and absorption	4.16E-06
	hsa00980	Metabolism of xenobiotics by cytochrome P450	2.51E-04
BRCA	hsa04512	ECM-receptor interaction	1.30E-05
	hsa04510	Focal adhesion	1.05E-04
CESC	**hsa05205**	**Proteoglycans in cancer[14]**	3.10E-04
DLBC	**hsa05330**	**Allograft rejection[8]**	3.91E-17
	hsa05332	**Graft-versus-host disease[3]**	1.15E-16
	hsa04940	Type I diabetes mellitus	5.47E-16
	hsa05320	Autoimmune thyroid disease	3.78E-14
	hsa05150	Staphylococcus aureus infection	7.73E-14
	hsa04672	Intestinal immune network for IgA production	1.06E-13
	hsa05416	Viral myocarditis	2.12E-13
	hsa04612	Antigen processing and presentation	2.70E-13
	hsa04145	Phagosome	1.31E-12
	hsa05310	Asthma 13	1.53E-11
	hsa05321	Inflammatory bowel disease (IBD)	2.23E-11
	hsa05166	HTLV-I infection	4.38E-11
	hsa05323	Rheumatoid arthritis	4.57E-11
	hsa05140	Leishmaniasis	1.22E-10
	hsa05168	Herpes simplex infection	2.50E-09
	hsa04514	**Cell adhesion molecules (CAMs)**	2.36E-08
	hsa04064	**NF-kappa B signaling pathway**	1.75E-07
	hsa05152	Tuberculosis	1.77E-07
	hsa05340	**Primary immunodeficiency**	3.63E-07
	hsa04060	Cytokine-cytokine receptor interaction	7.92E-07
	hsa05145	Toxoplasmosis	7.60E-06
	hsa05322	**Systemic lupus erythematosus**	3.30E-05
	hsa04062	**Chemokine signaling pathway**	9.90E-05
	hsa05164	Influenza A	1.60E-04
	hsa04662	**B cell receptor signaling pathway**	1.62E-04
	hsa04640	Hematopoietic cell lineage	1.69E-04
	hsa04666	**Fc gamma R-mediated phagocytosis**	7.20E-04
ESCA	hsa04512	ECM-receptor interaction	2.40E-06
	hsa04510	Focal adhesion	1.28E-05
	hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	2.97E-05
	hsa04974	Protein digestion and absorption	1.40E-04
HNSC	**hsa05414**	**Dilated cardiomyopathy[6]**	2.01E-06
	hsa04512	ECM-receptor interaction	2.78E-06
	hsa05410	Hypertrophic cardiomyopathy (HCM)	8.68E-06
	hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	2.47E-04
	hsa04261	Adrenergic signaling in cardiomyocytes	2.54E-04
	hsa04260	Cardiac muscle contraction	3.47E-04
LAML	**hsa04672**	**Intestinal immune network for IgA production**	1.27E-04
	hsa05140	**Leishmaniasis[20]**	3.11E-04
LIHC	hsa04610	Complement and coagulation cascades	6.85E-15
	hsa05150	Staphylococcus aureus infection	3.25E-05
	hsa00830	Retinol metabolism	7.71E-04
LUAD	hsa00512	Mucin type O-Glycan biosynthesis	4.57E-04
MESO	hsa04512	ECM-receptor interaction	7.29E-10
	hsa04610	Complement and coagulation cascades	3.37E-09
	hsa05150	**Staphylococcus aureus infection**	2.42E-08
	hsa04974	Protein digestion and absorption	5.75E-07
	hsa04510	Focal adhesion	8.13E-07
	hsa04151	**PI3K-Akt signaling pathway[15]**	9.60E-05
	hsa04145	Phagosome	1.73E-04
	hsa04611	**Platelet activation[15]**	7.07E-04
PAAD	hsa04974	Protein digestion and absorption	3.48E-09
	hsa04950	Maturity onset diabetes of the young	8.01E-07
	hsa04512	ECM-receptor interaction	1.38E-05
PCPG	hsa04721	Synaptic vesicle cycle	4.53E-05
PRAD	**hsa04270**	**Vascular smooth muscle contraction**	2.69E-04
SARC	hsa04512	ECM-receptor interaction	5.42E-11
	hsa04974	Protein digestion and absorption	7.52E-10
	hsa04510	Focal adhesion	1.32E-09
	hsa05146	**Amoebiasis**	2.89E-05
	hsa04151	PI3K-Akt signaling pathway	2.55E-04
SKCM	hsa04512	ECM-receptor interaction	3.08E-08
	hsa04510	Focal adhesion	6.44E-08
	hsa04974	Protein digestion and absorption	3.02E-07
	hsa05222	Small cell lung cancer	7.66E-05
	hsa05200	Pathways in cancer	8.55E-05
	hsa04151	PI3K-Akt signaling pathway	1.29E-04
STAD	**hsa04950**	**Maturity onset diabetes of the young[19]**	9.03E-04
TGCT	hsa04974	Protein digestion and absorption	4.96E-10
	hsa04512	ECM-receptor interaction	1.74E-05
THCA	hsa04512	ECM-receptor interaction	1.15E-05
	hsa04510	Focal adhesion	3.40E-04
	hsa04918	Thyroid hormone synthesis	5.91E-04
THYM	hsa04640	Hematopoietic cell lineage	6.43E-09
	hsa05340	**Primary immunodeficiency[18]**	2.67E-04
	hsa05162	Measles	9.46E-04
UCEC	**hsa04530**	**Tight junction**	5.60E-05
UCS	hsa04512	ECM-receptor interaction	1.02E-04
	hsa04974	Protein digestion and absorption	1.12E-04
	hsa04510	Focal adhesion	5.05E-04
KIRP	hsa04512	ECM-receptor interaction	5.93E-05
	hsa00040	**Pentose and glucuronate interconversions**	2.43E-04
	hsa00053	Ascorbate and aldarate metabolism	4.92E-04
	hsa00140	**Steroid hormone biosynthesis**	5.79E-04
KIRC	**hsa04610**	**Complement and coagulation cascades**	1.21E-06