Research Article
Journal of Intelligent Systems: Theory and Applications 4(2) 2021: 136-141
DOI: 10.38016/jista.946954

Convolutional Neural Network Approach to Predict Tumor Samples Using Gene Expression Data

Büşra Nur Darendeli [1], Alper Yılmaz [2*]

[1,2] Yıldız Technical University, Faculty of Chemical and Metallurgical Engineering, Department of Bioengineering, İstanbul/Turkey

bndarendeli@gmail.com, alyilmaz@yildiz.edu.tr

Abstract

Cancer is threatening millions of people each year and its early diagnosis is still a challenging task. Early diagnosis is one of the major ways to tackle the disease and lower the mortality rate. Advancements in deep learning approaches and the availability of biological data offer applications that can facilitate the diagnosis and characterization of cancer. Here, we aimed to provide a new perspective of cancer diagnosis using a deep learning approach on gene expression data.

In this study, RNA-Seq data of approximately 30 different types of cancer patients the Cancer Genome Atlas (TCGA) study, and normal tissue RNA-Seq data from GTEx were used. The input data for the training was transformed to RGB format and the training was carried out with a Convolutional Neural Network (CNN). The trained algorithm is able to predict cancer with 97% accuracy, using gene expression data. In conclusion, our study shows that the deep learning approach and biological data have a huge potential in the diagnosis and identification of tumor samples.

Keywords: cancer, CNN, gene expression, RNA-Seq, TCGA

Gen İfade Verilerinde Konvolusyonel Sinir Ağı Kullanılarak Tümör Örneklerinin Tahmini

Öz

Kanser her yıl milyonlarca insanı tehdit eden, erken teşhisi hala mümkün olmayan yaygın bir hastalıktır. Erken teşhis, kanserle baş etmenin ve ölüm oranını düşürmenin en önemli yollarından biridir. Derin öğrenme yaklaşımlarındaki gelişmeler ve biyolojik verilerdeki artış, kanserin teşhisini ve karakterizasyonunu kolaylaştırabilecek uygulamalar sunmaktadır. Bu çalışmada, gen ifade verilerini kullanarak derin öğrenme yaklaşımı ile kanser teşhisine yeni bir bakış açısı sağlamayı amaçladık.

30 farklı kanser çeşidine ait RNA-Seq verisi Kanser Genom Atlası (TCGA) adlı kaynaktan normal dokuların RNA-Seq verileri GTEx adlı kaynaktan temin edilip model eğitiminde kullanılmıştır. Gen ifade verileri RGB formatına dönüştürülüp Konvolusyonel Sinir Ağı (CNN) eğitimi için kullanıldı. Eğitilen model, gen ifade verilerine dayanarak kanseri %97 doğrulukla tahmin edebilmektedir. Sonuç olarak çalışmamız, derin öğrenme yaklaşımının ve biyolojik verilerin tümör örneklerinin tanısında büyük bir potansiyele sahip olduğunu göstermektedir.

Anahtar Kelimeler: CNN, Gen İfadesi, Kanser, RNA-Seq, TCGA.

1. Introduction

The deep learning approach has emerged by designing computer models that can perform the learning process as a result of interconnected layers based on the human brain, such as neurons. As a result of the development of data science and especially the rapid increase in biological data in the last decade, neural networks have begun to play important roles in the interpretation of biological data for the diagnosis and treatment of diseases (Esteva *et al.* 2019). Cancer, one

* Corresponding Author.
E-mail: alyilmaz@yildiz.edu.tr

Recieved : 02 June 2021
Revision : 04 July 2021
Accepted : 04 July 2021

of the biggest health problems in the world, is one of the diseases in which deep learning approaches were widely applied.

Since cancer is a disease with high genomic heterogeneity and phenotypic plasticity, its diagnosis and treatment are quite challenging (Persi *et al.*2020). Thanks to the developments in medical technologies, various forms of medical data are abundant for cancer patients and these data are suitable for deep learning-based approaches for diagnosis or treatment of cancer.

Image-based methods, such as radiology and histopathology, are commonly used for cancer diagnosis thus image-based deep learning approaches have thrived in supervised learning applications of cancer prediction. CT, MRI, histopathology data have been used in deep learning algorithms for the diagnosis of many types of cancers, including breast cancer (Zuluaga-Gomez *et al.* 2020; Gour *et al.* 2020; Zhang *et al.*; Hu *et al.* 2020; Bejnordi *et al.* 2018; Couture, *et al.* 2018), prostate cancer (Swiderska-Chadaj *et al.* 2020; Hartenstein *et al.* 2020; Duran-Lopez *et al.* 2020; Yoo *et al.* 2019; Nagpal *et al.* 2019; Arvaniti *et al.* 2018), lung cancer (Kanavati *et al.* 2020; Lai *et al.* 2020; Parnian *et al.* 2020), colon cancer (Jiang *et al.* 2020), head and neck cancer (Fontaine *et al.* 2020), and skin cancer (Tschandl *et al.* 2020, Esteva *et al.* 2017). These image-based studies facilitated clinical decision making especially in the detection of tumors in the early stages.

In addition to the image-based approaches, biological data such as gene expression (Dolezal *et al.* 2020) and gene mutations (Jiao *et al.* 2020) have also been used for training deep learning models to diagnose cancer. Gene expression data and deep learning approaches are integrated to tackle various challenges such as estimation of survival times of individuals with cancer (Ramirez *et al.* 2021), determination of biomarker genes (Xie *et al.* 2021), assurance of effective therapeutics for cancer treatment (Zeng *et al.* 2021), classification of cancer subtypes (Binder *et al.* 2021, Galili *et al.* 2021, Ahn *et al.* 2018). Ahn *et al.* developed a deep learning algorithm using publicly available gene expression databases to classify the samples as normal or tumor and high predictive scores were obtained. All of these studies show that by using gene expression data and deep learning approaches together, critical information will be revealed about the mechanism of cancer.

In our study, The Cancer Genome Atlas (TCGA) dataset with RNA-Seq data of approximately 30 different types of cancer patients and a dataset obtained by curation of GTEx data including RNA-Seq analysis of normal tissues was used. The input data for the training was converted to RGB format and the training was carried out with the CNN algorithm. The trained algorithm can predict cancer and normal patients with 97% accuracy, based on gene expression data. Our results suggest that gene expression data have the potential to make inferences on cancer by mapping gene expression content to RGB space.

2. Methods

2.1 Dataset Preparation

Data was downloaded from the UCSC Xena platform (UCSC Xena), which includes RNA-Seq data from various resources including, TCGA and GTEx. Label distribution of selected datasets is shown in Table 1.

Table 1. Distribution of training dataset labels.

Datasets	*Normal*	*Tumor*
TCGA	727	9750
GTEx	7429	0

Data labels (Normal, Tumor) have been extracted from phenotype information of selected samples. Gene IDs were converted from Entrez ID to ENSEMBL IDs using the BioMart online tool (BioMart).

The differentially expressed gene list (LINCS Harmonizome)(Rouillard et al. 2016) was used to select 1024 genes that show the highest up-regulation or down-regulation count throughout the whole dataset. Expression data for selected genes have been using as input for training.

2.2 Conversion of Inputs to Images

Gene expression values have converted into (R, G, B) format before the training step. RGB values are obtained by converting gene expression value into 24-bit long binary and then using the first 8 bits for R (red), second 8 bits for G (green), and third 8 bits for B (blue) (Figure 1). For each sample, a 32x32x3 3D Numpy array was prepared.

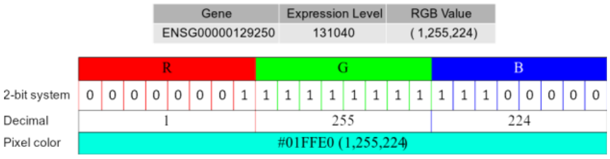


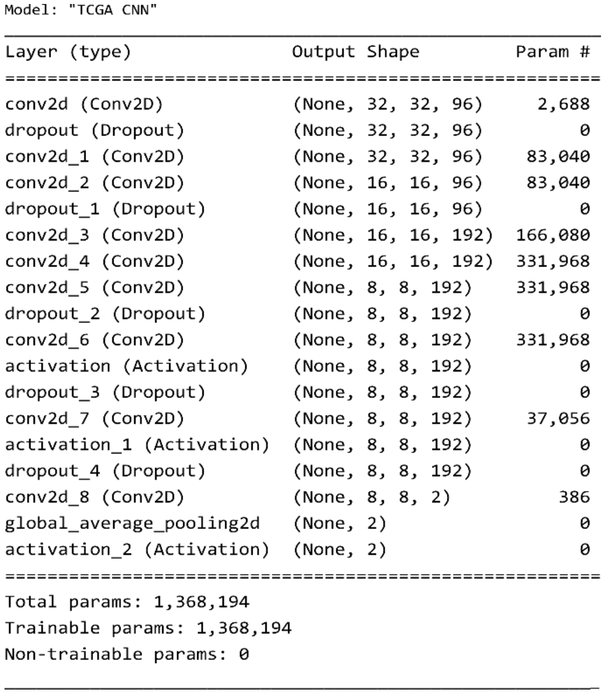
Figure 1. Conversion of gene expression value to RGB format.

2.3 CNN Architecture

The CNN architecture shown in Table 2 has been using for training. The architecture includes eight convolution layers, four dropout layers, one global average pooling layer. Each convolution layer consists of 3x3 kernels.

ReLU has been using as an activation function and, to overcome overfitting, dropout rates of 0.2 or 0.5 used. The final layer has a Sigmoid as an activation function.

Table 2. CNN Architecture



```
Model: "TCGA CNN"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 32, 32, 96)        2,688
dropout (Dropout)            (None, 32, 32, 96)        0
conv2d_1 (Conv2D)            (None, 32, 32, 96)        83,040
conv2d_2 (Conv2D)            (None, 16, 16, 96)        83,040
dropout_1 (Dropout)          (None, 16, 16, 96)        0
conv2d_3 (Conv2D)            (None, 16, 16, 192)       166,080
conv2d_4 (Conv2D)            (None, 16, 16, 192)       331,968
conv2d_5 (Conv2D)            (None, 8, 8, 192)         331,968
dropout_2 (Dropout)          (None, 8, 8, 192)         0
conv2d_6 (Conv2D)            (None, 8, 8, 192)         331,968
activation (Activation)      (None, 8, 8, 192)         0
dropout_3 (Dropout)          (None, 8, 8, 192)         0
conv2d_7 (Conv2D)            (None, 8, 8, 192)         37,056
activation_1 (Activation)    (None, 8, 8, 192)         0
dropout_4 (Dropout)          (None, 8, 8, 192)         0
conv2d_8 (Conv2D)            (None, 8, 8, 2)           386
global_average_pooling2d     (None, 2)                 0
activation_2 (Activation)    (None, 2)                 0
=================================================================
Total params: 1,368,194
Trainable params: 1,368,194
Non-trainable params: 0
_________________________________________________________________
```

3. Results

3.1. Retrieved Input Images

Since gene expression data have been converted into RGB format, visualizing the expression layout for any sample as possible. In Figure 2, sample images for Normal and Tumor samples are presented. The images do not reveal any apparent pattern for the naked eye. However, convolutional layers are able to pick regions or patterns formed by neighboring pixels so gene expression data was passed through convolution layers. Please note that gene expression data was converted into RGB format but they are not saved as images before training. The training was performed on a 32x32x3 3D multidimensional array for each sample.

3.2. CNN Training

The deep learning architecture shown in Table 2 has been using for the training of 17,906 samples having evenly distributed normal and tumor labels. Samples were split into Train: Test with 80:20 ratio. After 40 epochs the accuracy has reached 97.7%. The accuracy and loss plots of the test and training samples are shown in Figure 3.

3.3 Performance Measurement

Figure 4 shows the ROC curve of the model. The AUC value of our model was found to be 0.97. Additional performance measures were calculated from the confusion matrix generated by test sample predictions. Our model had 98% precision and 98% recall for tumor prediction (Table 3).

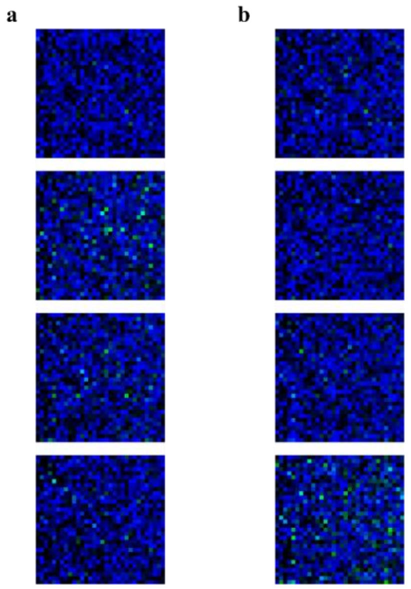


Figure 2. Visualization of gene expression data as image. 4 sample images from (a) Normal tissue data and (b) Tumor tissue data generated by converting gene expression levels of 1024 selected genes using RGB mapping.

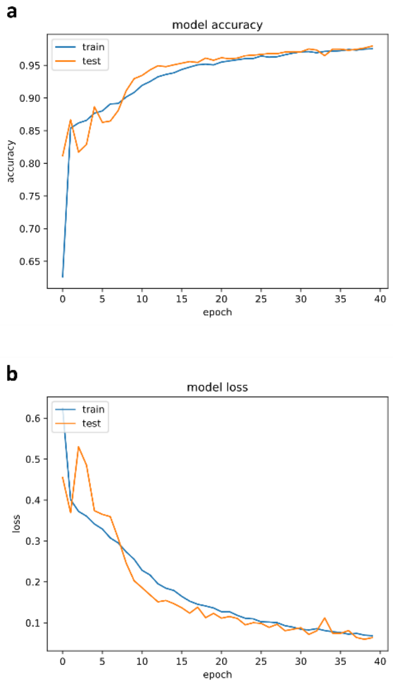


Figure 3. Model accuracy (a) and loss (b) plots.

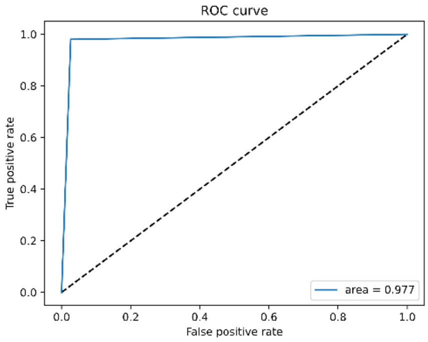


Figure 4. The ROC curve of CNN model test predictions for tumor and normal classification.

Table 3. Performance measurements based on confusion matrix

	Tumor Prediction
Accuracy	0.98
Precision	0.97
Recall	0.98
F1-Score	0.98

In literature, several different approaches use gene expression data to classify tumor and normal samples ranging from simpler machine learning approaches to complex deep learning networks. These approaches usually start with pre-processing the gene expression data with an irreversible manipulation (normalization) and even mapping data points to a different domain (PCA, t-SNE, etc.). Our method involves a minimal and reversible change to gene expression data. The RGB mapping is reversible and does not require normalization or any dimensional reduction techniques. Table 4 compares our approach with several different approaches both in pre-processing and classification steps. Although Elbashir et al. study (Normalization + CNN) has the highest accuracy, the sample used in their study is problematic so our approach has better results overall.

Please note that Elbashir et al uses a smaller and unbalanced TCGA dataset (only Breast Cancer dataset, containing 113 Normal, 1095 Tumor samples). Their accuracy starts from 91% and reaches 98.7% and, due to dominating the number of tumor samples, their model has a tendency to pick "tumor" as a label irrespective of the sample being predicted, explaining their very low sensitivity and full precision scores. In our case, our dataset is larger and balanced (8156 Normal vs. 9750 Tumor) and, our accuracy starts from 54% and then reaches 97.7%.

4. Conclusions

Due to its complex biological microenvironment, cancer has many difficulties in diagnosis and treatment. The difficulties caused by this complexity can be overcome with ever-increasing RNA-Seq data. The vast number of expression data sets combined with deep learning models have the potential to help diagnose cancer cases.

In this study, we proposed an approach to process gene expression in a reversible manner that does not require normalization. After RGB mapping of expression data, the processed data can be treated as image data and be subject to convolutional neural network learning.

Since our approach retains each pixel as an individual gene, segmentation analysis which reveals important pixels has the potential to reveal important genes for cancer development. Moreover, the strength of RGB mapping should be tested for not only tumor prediction but also tumor stage prediction.

Table 4. Comparison of our model with other studies. SVM; support vector machine, t-SNE; t-distrubuted stochastic neighbor embedding.

Expression Preprocessing	**Classification**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F-measure**	**References**
RGB mapping	CNN	97,73%	97,66%	97,80%	98,00%	0,975	Our method
Normalization	CNN	98,76%	91,43%	100,00%	100,00%	0,955	Elbashir et al.
Normalization	Stacked Denoising	94,78%	94,04%	97,50%	97,20%		Danaee et al.
Normalization	AlexNet	96,69%	96,89%	94,12%	99,54%	0,955	Elbashir et al
t-SNE	SVM	100,00%	100,00%	51,00%	95,96%	0,97	Elbashir et al

References

Ahmed, O., & Brifcani, A. (2019, April). Gene Expression Classification Based on Deep Learning. In 2019 4th Scientific International Conference Najaf (SICN) (pp. 145-149). IEEE.

Ahn, T., Goo, T., Lee, C. H., Kim, S., Han, K., Park, S., & Park, T., 2018. Deep learning-based identification of cancer or normal tissue using gene expression data.

In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1748-1752). IEEE.

Arvaniti, E.,Fricker, K.S.,Moret, M.,Rupp, N.,Hermanns, T.,Fankhauser, C.,Wey, N., Wild, P.J.,Rueschoff, J.H. and Claassen, M., 2018. Automated Gleason grading of prostate cancer tissue microarras via deep learning. Scientific reports,8(1), pp.1-11.

Bejnordi, B.E., Mullooly, M., Pfeiffer, R.M., Fan, S.,Vacek, P.M., Weaver, D.L., Herschorn, S., Brinton, L.A., van Ginneken, B., Karssemeijer, N. and Beck, A.H., 2018. Using deep convolutional neural networks to identify and classify tumor associated stroma in diagnostic breast biopsies. Modern Pathology, 31(10), pp.1502-1512.

Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., ... & Klauschen, F. (2021). Morphological and molecular breast cancer profiling through explainable machine learning. Nature Machine Intelligence, 1-12.

Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J.S., Perou, C.M., Troester, M.A. and Niethammer, M., 2018. Image analysis with deep learning to predict breast cancer grade, Erstatus, histologic subtype, and intrinsic subtype. NPJ breast cancer, 4(1), pp.1-8.

Danaee, P., Ghaeini, R., & Hendrix, D. A. (2017). A deep learning approach for cancer detection and relevant gene identification. In Pacific symposium on biocomputing 2017 (pp. 219-229).

Dolezal, J.M., Trzcinska, A., Liao, C.Y., Kochanny, S., Blair, E., Agrawal, N., Keutgen, X.M., Angelos, P., Cipriani, N.A. and Pearson, A.T., 2020. Deep learning prediction of BRAF- RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. Modern Pathology, pp.1-13.

Duran-Lopez, L., Dominguez-Morales, J.P., Conde-Martin, A.F., Vicente-Diaz, S. and Linares- Barranco, A., 2020. PROMETEO: A CNN-Based Computer-Aided Diagnosis System for WSI Prostate Cancer Detection. IEEE Access, 8, pp.128613-128628.

Elbashir, M. K., Ezz, M., Mohammed, M., & Saloum, S. S. (2019). Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data. IEEE Access, 7, 185338-185348

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), pp.115-118.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J., 2019. Aguide to deep learning in healthcare. Nature medicine, 25(1), pp.24-29.

Fontaine, P., Acosta, O., Castelli, J., De Crevoisier, R., Müller, H. and Depeursinge, A., 2020. The importance of feature aggregation in radiomics: a head and neck cancer study. Scientific Reports, 10(1), pp.1-11.

Galili, B., Tekpli, X., Kristensen, V. N., & Yakhini, Z., 2021. Efficient gene expression signature for a breast cancer immuno-subtype. Plos one, 16(1), e0245215.

Gour, M., Jain, S. and SunilKumar, T., 2020. Residual learning based CNN for breast cancer histopathological image classification. International Journal of Imaging Systems and Technology.

Hartenstein, A., Lübbe, F., Baur, A.D., Rudolph, M.M., Furth, C., Brenner, W., Amthauer, H., Hamm, B., Makowski, M. and Penzkofer, T., 2020. Prostate Cancer Nodal Staging: Using Deep Learning to Predict 68 Ga-PSMA-Positivity from CT Imaging Alone. Scientific Reports, 10(1), pp.1-11.

Hu, Q., Whitney, H.M. and Giger, M.L., 2020. Adeep learning methodology for improved breast cancer diagnosis using multiparametric MRI. Scientific Reports, 10(1), pp.1-11.

Jiang, D., Liao, J., Duan, H., Wu, Q., Owen, G. Shu, C., Chen, L., He, Y., Wu, Z., He, D. and Zhang, W., 2020. A machine learning-based prognostic predictor for stage III colon cancer. Scientific reports, 10(1), pp.1-9.

Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., De Ridder, J., van Herpen, C., Lolkema, M.P., Steeghs, N. and Getz, G., 2020. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nature communications, 11(1), pp.1-12.

Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O. and Tsuneki, M., 2020. Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific Reports, 10(1), pp.1-11.

Li, Z., Zou, D., Tang, J., Zhang, Z., Sun, M., & Jin, H., 2019. A comparative study of deep learning-based vulnerability detection system. IEEE Access, 7, 103184-103197.

Lai, Y.H., Chen, W.N., Hsu, T.C., Lin, C., Tsao, Y. and Wu, S., 2020. overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Scientific reports, 10(1), pp.1-11.

Marra, F., Gragnaniello, D., & Verdoliva, L., 2018. On the vulnerability of deep learning to adversarial attacks for camera model identification. Signal Processing: Image Communication, 65, 240-248.

Mencattini, A., Di Giuseppe, D., Comes, M.C., Casti, P., Corsi, F., Bertani, F.R., Ghibelli, L., Businaro, L., Di Natale, C., Parrini, M.C. and Martinelli, E., 2020. Discovering the hidden messages within cell trajectories using a deep learning approach for in vitro evaluation of cancer drug treatments. Scientific reports, 10(1), pp.1-11.

Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L. Mohtashamian, A., Wren, J.H. and Corrado, G.S., 2019. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ digital medicine, 2(1), pp.1-10.

Parnian, A., Arash, M., Tyrrell, P.N., Cheung, P., Ahmed, S., Plataniotis, K.N., Nguyen, E.T. and Anastasia, O., 2020. DRTOP: deep learning-based radiomics for the time-to-event outcome prediction in lung cancer. Scientific Reports (Nature Publisher Group), 10(1).

Persi, E., Wolf, Y.I., Horn, D., Ruppin, E., Demichelis, F., Gatenby, R.A., Gillies, R.J. and Koonin, E.V., 2020. Mutation–selection balance and compensatory mechanisms in tumour evolution. Nature Reviews Genetics, pp.1-12.

Ramirez, R., Chiu, Y. C., Zhang, S., Ramirez, J., Chen, Y., Huang, Y., & Jin, Y. F., 2021. Prediction and interpretation of cancer survival using graph convolution neural networks. Methods.

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M.G., & Ma'ayan, A., 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database, 2016.

Shon, H. S., Yi, Y., Kim, K. O., Cha, E. J., & Kim, K. A. (2019). Classification of stomach cancer gene expression data using CNN algorithm of deep learning. Journal of Biomedical and Translational Research (JBTR), 20(1), 15-20.

Sinha, S., & Saranya, S. S., 2021. One Pixel Attack for Fooling Neural Networks. Annals of the Romanian Society for Cell Biology, 8405-8412.

Su, J., Vargas, D. V., & Sakurai, K., 2019. Attacking convolutional neural network using differential evolution. IPSJ Transactions on Computer Vision and Applications, 11(1), 1-16.

Swiderska-Chadaj, Z., de Bel, T., Blanchet, L., Baidoshvili, A., Vossen, D., van der Laak, J. and Litjens, G., 2020. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. Scientific Reports, 10(1), pp.1-14.

Tran, N.H., Qiao, R., Xin, L., Chen, X., Shan, B. and Li, M., 2020. Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. Nature Machine Intelligence, pp.1-8.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J. and Paoli, J., 2020. Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8), pp.1229-1234.

Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H., 2021. Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational oncology, 14(1), 100907.

Yoo, S., Gujrathi, I., Haider, M.A. an Khalvati, F., 2019. prostate cancer Detection using Deep convolutional neural networks. Scientific Reports, 9.

Zeng, B., Glicksberg, B. S., Newbury, P., Chekalin, E., Xing, J., Liu, K., ... & Chen, B., 2021. OCTAD: an open workspace for virtually screening therapeutics targeting precise cancer patient groups using gene expression features. Nature Protocols, 16(2), 728-753.

Zhang, Y., Chan, S., Park, V.Y., Chang, K.T., Mehta, S., Kim, M.J., Combs, F.J., Chang, P., Chow, D., Parajuli, R. and Mehta, R.S., 2020. Automatic Detection and Segmentation of Breast Cancer on MRI Using Mask R-CNN Trained on Non–Fat-Sat Images and Tested on Fat-Sat Images. Academic Radiology.

Zuluaga-Gomez, J., Al Masry, Z., Benaggoune, K., Meraghni, S. and Zerhouni, N., 2020. A CNN-based methodology for breast cancer diagnosis using thermal images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, pp.1-15.