A DEEPFAKE-BASED APPROACH FOR GENERATING CANCEROUS GENE EXPRESSION DATA

Al-Zahraa Eid, Amany Bahaa El-Din, Esraa Sayed, Galal Hossam

Abstract

Radiogenomics is a newly emerging field that integrates genomic (gene activity) with radiomic (medical imaging) data for elucidating the associations between gene expression data and imaging phenotypes, especially in cancer. The shortage in the number of the available paired data (genomic and imaging) for the same cohort of patients hinders the research of mapping the gene expression features with the imaging features and the comprehensive understanding of the underlying mechanisms by which cancer work. One way to mitigate these concerns is by computationally generating numerous amounts of realistic data of either/both types for widely spread carcinomas. We propose a deep learning approach based on Wasserstein generative adversarial network with gradient penalty (WGAN-GP) to generate gene expression data that are hardly differentiable from realistic data with application on Glioblastoma (GBM) and Kidney Renal Clear Cell Carcinoma (KIRC) genetic data sets. To select cancer-related genes for training, we used the disease pathway database and the inference of the differentially expressed genes (DEGs). By training the model on the two cancer types, it achieved generator score of 0.91 for GBM and 0.81 for KIRC. Testing the realism of the generated data of GBM and KIRC showed realism score of 0.9 and 0.78 respectively. We visualized the pairwise distances between the gene expressions using the distance matrix, the dendrograms, and the plotting of the distribution of the real and generated gene expression profiles. Our results indicate that the computational generation of the gene expression data is a possible method that could be used to increase the number and size of the available gene expression dataset.

Keywords: Radiogenomics,WGAN-GP

I. Introduction

Over the last decade, radiogenomics field has grown rapidly, demonstrating enormous potential for developing non-invasive prognostic and diagnostic approaches, as well as identifying biomarkers for cancer treatment, by combining quantitative imaging features with tumor phenotyping and genomic signature [1]. As the healthcare researchers need large datasets for their researches in Radiogenomics, which can be provided by introducing a deep learning-based approach that can augment realistic imaging and genetic data. Because of having the main shortage in the number of available gene expression datasets, our main concern is to generate realistic genetic data for widely spread carcinomas. We propose a deep learning-based approach to generate gene expression data that are hardly differentiable from realistic data, and apply it on Glioblastoma (GBM) and Kidney Renal Clear Cell Carcinoma (KIRC) genetic data set. The proposed approach uses powerful deep learning tools, depending mainly on the GANs (Generative Adversarial Networks). GANs architecture has two basic elements: the generator network and the discriminator network. The generator network generates fake data and the discriminator differentiate between the fake data and the realistic data and send its feedback to the generator network which uses this feedback to improve the quality of the generated data.

II. Materials and methods

A. The generative adversarial model:

Our method is built on WGAN-GP proposed by Arjovsky et al. [2], which is an alternative to the traditional GAN. It the same architecture, the generator G and critic D. The generator G has three inputs, which are noise vector *z*, sample covariates *r* and *q*, and it produces a vector $\hat{x}$ of synthetic expression values as an output. The critic has three inputs which are real gene expression sample x or a synthetic sample $\hat{x}$, in addition to sample covariates *r* and *q*, and its main goal is trying to distinguish whether the input sample is real or fake. For the two components, we use word embedding technique to model the sample covariates, a unique feature that allows learning distributed, dense

representations for the different tissue-types, and we use it generally for all categorical covariates $q \in N^c$.

The cost functions used for updating the parameters for each network are:

- Generator's loss function:

$$- E_{\hat{x}\sim P_g}[D(\hat{x}, r, q)]$$

- Critci's loss function:

$$E_{\hat{x}\sim P_g}[D(\hat{x}, r, q)] - E_{x\sim P_r}[D(x, r, q)] - \lambda E_{\bar{x}\sim P}[||\nabla_{\bar{x}} D(\bar{x}, r, q)||2 - 1]^2$$

B. *The datasets:*

- GBM dataset:

Glioblastoma (GBM) is one of the most common brain tumors, which has the characteristics of high morbidity, high recurrence, high mortality and low cure rate [3]. We downloaded mRNAseq_693 dataset for Glioblastoma (GBM) [4,5] from the Chinese Glioma Genome Atlas (CGGA), which is a user-friendly data portal for the storage and interactive exploration of genomic data from Chinese glioma patients [6]. The dataset contains 693 samples and 23,987 genes. But we selected 71 genes for all samples according to the GBM disease pathways. We also included the tissue type (Primary, Recurrent) and the sex as a categorical covariate, and the age as numerical covariate.

- KIRC dataset:

We worked with the TCGA-KIRC data set for kidney cancer from The Cancer Genome Atlas (TCGA), It has 243 samples and 16383 genes. To work on these data, we should select the DEGs which refers to the differentially expressed genes between the normal and cancerous data.

C. *The data preprocessing:*

- Gene selection:
  - The differentially expressed genes (DEGS):

Differentially expressed genes (DEGs) are the genes whose gene expression levels differ from one condition to another [7].

Steps for getting the DEGs:

1. Data filteration
2. Hypothesis testing using Wilcoxon signed rank test which is a nonparametric version of the paired Student t-test followed by multiple hypothesis testing using the Bonferroni method which is a correction method that redefines the significance level to be α/m , where m is the number of comparisons. select the genes with corrected p-values is less than or equal to the original significance level.
3. Fold Change: Fold Change (FC) is a quantitative measure for the change in the expression levels under two different conditions. It can be computed using the following equation:

$$\text{FC} = \frac{\text{the average GE level of the healthy samples}}{\text{the average GE level of the diseased samples}}$$

Then get log2(FC), select the genes that has |log2(FC)| $\geq$ log2(T), where T is a threshold for significance of the difference.

4. The intersection of DEGs of the Hypothesis Testing method with the DEGs of the FC method should make the disease list of genes.
5. Visualize DEGs using volcano plot which is a type of scatterplot that shows statistical significance (P value) versus magnitude of change (fold change).

- The disease pathway:

The disease pathway is a series of actions among molecules in a cell that leads to a certain disease or a change in the cell. We used the pathways introduced in the KEGG pathway maps for human diseases [8].

- Data normalization:

To make the learning rate easier, we standardize the expression values to keep them within a reasonable range.

$$\boldsymbol{Z_{i,j}} = \frac{\boldsymbol{X_{i,j}} - \boldsymbol{\mu_j}}{\boldsymbol{\sigma_j}}$$

Where:

$$\boldsymbol{\mu_j} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{X_{i,j}} \quad \text{and} \quad \boldsymbol{\sigma_j} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{X_{i,j}} - \boldsymbol{\mu_j})^2}$$

D. *Evaluating generated gene expression:*

To evaluate our synthetic data. We calculated a similarity coefficient using Pearson's correlation coefficient. Let $\boldsymbol{A}$ be a n x n symmetric matrix that contains all of the pairwise distances between all the genes. To determine how well this matrix preserves pairwise distances when compared to another n x n distance matrix $\boldsymbol{B}$.

$$\gamma(A,B) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{A_{i,j} - \mu(A)}{\sigma(A)}\right)\left(\frac{B_{i,j} - \mu(B)}{\sigma(B)}\right)$$

Where:

$$\mu(G) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} G_{i,j}$$

$$\sigma(G) = \sqrt{\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (G_{i,j} - \mu(G))^2}$$

$\gamma(A,B)$ computes the Pearson's correlation between the elements in the upper diagonal of matrices ***A***&***B***.

III. Results

A. Model Evaluation:

We trained our GAN on the GBM mRNAseq_693 dataset from CGGA and we trained once more on the TCGA-KIRC dataset, and the following table illustrates the training details and accuracy.

Table 1 The different scores of the model with different datasets

Cancer type	No. of samples	No. of genes	Method for gene selection	Model score	Evaluation (γ)
GBM	691	71	Disease pathway	93%	0.9
KIRC	243	46	Disease pathway	81%	0.78
KIRC	243	4197	DEGs	62%	0.53

B. Analyzing generated GBM gene expression data:

We calculated the distance matrix for the real and the synthetic gene expressions for the 12 most frewuent genes.

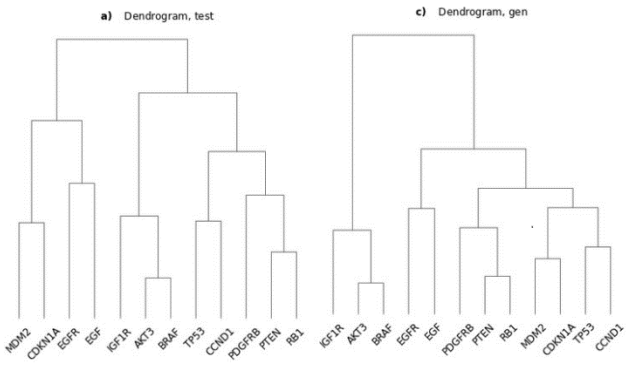

Figure 1 Dendrograms for a subset of 12 GBM most frequently mutant genes.

C. Visualizing the DEGs for KIRC dataset:

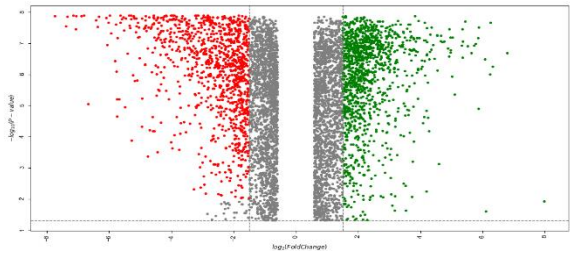

Figure 2 Volcano plot for visualizing DEGs for KIRC dataset

D. Analyzing generated KIRC gene expression data:

We calculated the distance matrix for the real and the synthetic gene expressions for the 12 genes.

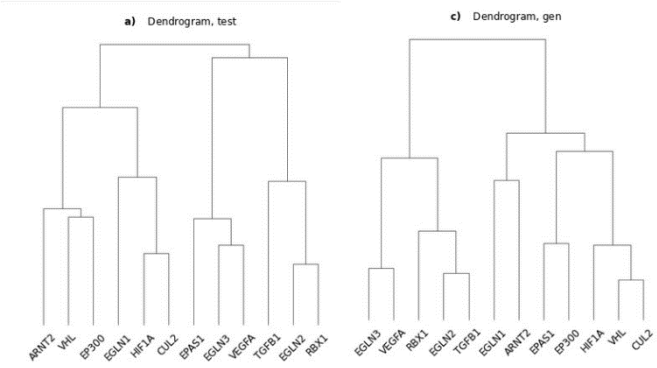

Figure 3 Dendrograms for a subset of first 12 KIRC disease pathway

IV. Dicussion

A. Datasets and preprocessing:

we used the dataset from the CGGA which has sufficient number of samples. To get the DEGs for the KIRC data, we made data filtration for the TCGA-KIRC data, then we calculated the corrected p-values of the genes using Bonferroni correction method and compared them to a significance level equal to 0.05, we kept the genes with corrected p-values <= 0.05. These genes are the DEGs using the hypothesis testing approach. We then calculated the lod2(FC) and we used T=1.5, if $|\log2(FC)| \geq 1.5$, this gene is a DEG. We loop for all the genes and get the DEGs of the fold change method. This intersection between the DEGs of hypothesis test and fold change method results to 4197genes. Finally, we draw the volcano plot between the log2(FC) on the x axis and -log10(P-Value) on the y axis.

B. The generative adversarial model:

We used the WGAN-GP for building our model and it was a good choice for building the model, because it does not have problems with convergence like GAN, or generating poor samples like WGAN.

To check the performance of the model, we used K-Fold cross validation, which is a resampling procedure for evaluating the model on limited data samples. We applied this procedure on the GBM dataset with K = 5 folds.

After training the model on the datasets using the two different gene selection ways, we deduced that the training on the selected genes from the disease pathway genes outputs larger score than the training on the selected genes from the DEGs

V. Conclusion and Future work:

In this report, we introduced a deep learning-based approach to generate gene expression data that are hardly differentiable from realistic data, using the WGAN-GP. We trained the model on datasets GBM and KIRC separately after selecting the genes that we are interested in, then we obtained the model score at each case, and it was of 0.91 for GBM and 0.81 for KIRC.

The results illustrates that the model managed to generate realistic gene expression data.

We will try to obtain segmented data that contain information about the tumor labels, to achieve generation of fake imaging data, and we will try to obtain sufficient imaging and genetic data for the same subject, to achieve generating both the radiomic and the genomic data.

VI. Acknowledgments

We thank Dr. Ibrahim Youssef for his supportive supervision and helpful comments that light our way during working on this project.

VII. Refernces:

[1] Panayides AS, Pattichis MS, Leandrou S, Pitris C, Constantinidou A, Pattichis CS. Radiogenomics for precision medicine with a big data analytics perspective. IEEE journal of biomedical and health informatics. 2018 Dec 25;23(5):2063-79.

[2] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028. 2017 Mar 31.

[3] Ostrom QT, Gittleman H, Farah P, Ondracek A, Chen Y, Wolinsky Y, Stroup NE, Kruchko C, Barnholtz-Sloan JS. CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006-2010. Neuro-oncology. 2013 Nov 1;15(suppl_2):ii1-56.

[4] Liu X, Li Y, Qian Z, Sun Z, Xu K, Wang K, Liu S, Fan X, Li S, Zhang Z, Jiang T. A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. NeuroImage: Clinical. 2018 Jan 1;20:1070-7.

[5] Wang Y, Qian T, You G, Peng X, Chen C, You Y, Yao K, Wu C, Ma J, Sha Z, Wang S. Localizing seizure-susceptible brain regions associated with low-grade gliomas using voxel-based lesion-symptom mapping. Neuro-oncology. 2015 Feb 1;17(2):282-8.

[6] Zhao Z, Zhang KN, Wang Q, Li G, Zeng F, Zhang Y, Wu F, Chai R, Wang Z, Zhang C, Zhang W. Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data from Chinese gliomas. Genomics, proteomics & bioinformatics. 2021 Mar 2.

[7] Anjum A, Jaggi S, Varghese E, Lall S, Bhowmik A, Rai A. Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. J Comput Biol. 2016 Apr;23(4):239-47. doi: 10.1089/cmb.2015.0205. Epub 2016 Mar 7. PMID: 26949988; PMCID: PMC4827276.

[8] Kanehisa M. A database for post-genome analysis. Trends in genetics: TIG. 1997 Sep 1;13(9):375-6.