A DEEPFAKE-BASED APPROACH FOR GENERATING CANCEROUS GENE EXPRESSION DATA

By

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

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
BACHELOR OF SCIENCE
in
Systems and Biomedical Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

A DEEPFAKE-BASED APPROACH FOR GENERATING RADIOGENOMICS DATA IN CANCER

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

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
BACHELOR OF SCIENCE
in
Systems and Biomedical Engineering

Under the Supervision of

Dr. Ibrahim Youssef
Assistant Professor
Systems and Biomedical Engineering Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

Acknowledgments

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

We also thank Academy of scientific research and technology for funding our project.

Table of Contents

ACKNOWLEDGMENTS ..... **III**

TABLE OF CONTENTS ..... **IV**

LIST OF TABLES ..... **VI**

LIST OF FIGURES ..... **VII**

ABSTRACT ..... **VIII**

CHAPTER 1 : INTRODUCTION **1**

1.1. BACKGROUND: 1
1.2. MOTIVATION: 1
1.3. THE OVERALL OBJECTIVE: 1
1.4. SMART OBJECTIVES: 1
- Specific: 1
- Measurable: 2
- Achievable: 2
- Relevant: 2
- Time bound: 2

1.5. GANTT CHART: 2

CHAPTER 2 : LITERATURE REVIEW **3**

CHAPTER 3 : MATERIALS AND METHODS **6**

3.1. GENERATING GENE EXPRESSION DATA: 6
3.1.1. Introduction to the generative adversarial model: 6
3.1.2. Datasets: 7
3.1.2.1. GBM dataset: 7
3.1.2.2. KIRC dataset: 8
3.1.3. Data preprocessing: 8
3.1.3.1. Gene selection: 8
3.1.3.1.1. Getting the differentially expressed genes (DEGs): *8*
3.1.3.1.2. Getting the disease pathway genes: *9*
3.1.3.2. Data normalization: 9
3.1.4. The model architecture: 9
3.1.5. Evaluating generated gene expression data: 10
3.1.5.1. Correlation between real and generated data distance matrices: 10
3.1.5.2. Correlation between real and generated data distance dendrogram: 11
3.2. GENERATING MEDICAL IMAGES: 11
3.2.1. Generating abnormal brain tumor MRI images 11
3.2.2. Generate glioblastoma brain MRI images: 11

CHAPTER 4 : RESULTS **12**

4.1. MODEL EVALUATION 12
4.2. ANALYZING GENERATED GENE EXPRESSION DATA 12
4.2.1. Analyzing generated GBM gene expression data: 12

4.2.2. Visualizing the DEGs for KIRC data: 14
4.2.3. Analyzing generated KIRC gene expression data: 14
CHAPTER 5 DISCUSSION 16
5.1. GENE EXPRESSION DATA GENERATION: 16
5.1.1. Datasets and preprocessing: 16
5.1.2. The generative adversarial model: 16
5.2. MEDICAL IMAGES GENERATION: 18
CHAPTER 6 CONCLUSIONS AND FUTURE WORK 19
6.1. CONCLUSION: 19
6.2. FUTURE WORK: 19
REFERENCES 20

List of Tables

Table 1: Literature review 3
Table 2: The different scores of the model with different datasets 12

List of Figures

Figure 1: Gantt Chart....2
Figure 2: Representation of the generator and discriminator....6
Figure 3: When the synthetic data resembles the real data, the distance matrices Dx and Dz are similar, and the distance 1- $\gamma(D^X, D^Z)$ is close to 0.....10
Figure 4: The Distance Matrix for real data on the left side and the distance matrix for generated data on the right side for the GBM carcinoma....12
Figure 5: dendrograms for a subset of 12 GBM most frequently mutant genes. 13
Figure 6 : The distribution of the real and generated GBM gene expressions....13
Figure 7: Volcano plot for visualizing DEGs of KIRC data 14
Figure 8: the distance matrix for real data on the the left side and the distance matrix for the generated data on the right side for KIRC....14
Figure 9: dendrograms for a subset of first 12 mutant genes in the KIRC disease pathways....15
Figure 10: The distribution of the real and generated KIRC gene expressions....15
Figure 11: The discriminator in the traditional GAN results in vanishing gradients, while the critic in WGAN provides smooth gradient everywhere. 17

Abstract

Radiogenomics is a newly emerging field that integrates genomic (gene activity) with radiomic (medical imaging) data with the aim of 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, the comprehensive understanding of the underlying mechanisms by which cancer work, and the training of the cancer clinicians. 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 datasets.

Chapter 1 : Introduction

1.1. Background:

Radiogenomics is a newly emerging field that integrates genomic with radiomic (medical imaging) data with the aim of elucidating the associations between gene expression data and imaging phenotypes, especially in cancer. Over the last decade, this 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].

1.2. Motivation:

Radiogenomics is a promising field that has many valuable potentials such as: (1) shedding light on the underlying disease mechanisms; (2) advancing personalized medicine by enhancing precision of diagnosis, assessment of prognosis, estimation of survival, and prediction of treatment response; and (3) inferring the disease molecular background by non-invasive correlated imaging features.

1.3. The Overall Objective:

Radiogenomics is hindered by the expensive cost of the genetic screening tests, and the unavailability of numerous big data sets of paired imaging and genetic data for the same subjects that are crucial for training machine learning-based techniques for the analysis of radiogenomic studies, 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.

1.4. SMART objectives:

- **Specific:**

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.

- **Measurable:**

Our approach is measurable as we can evaluate its performance by defining its accuracy and measure how real the generated data are using the pearson's correlation.

- **Achievable:**

The approach is achievable as there are many literature reviews working on the same application using GAN as it is a powerful tool and has quickly received great interest.

- **Relevant:**

Generating radiogenomic data provide numerous big data sets of paired imaging and genetic data for the same subjects without ethical and privacy concerns. These datasets are crucial for training machine learning-based techniques for the analysis of radiogenomic studies, which has the opportunity to improve cancer diagnosis.

- **Time bound:**

Our time bound is the duration of the academic year (about 8 months).

1.5. Gantt Chart:

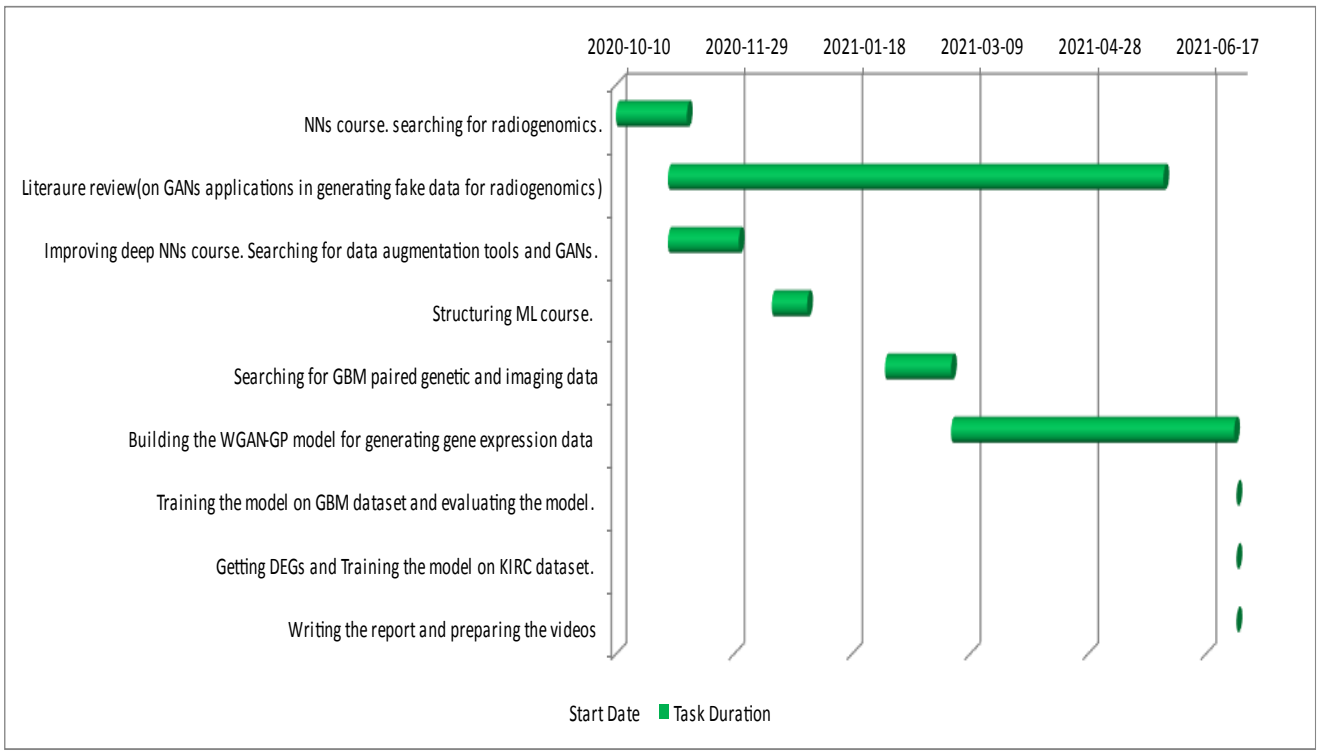

Figure 1: Gantt Chart

Chapter 2 : Literature Review

Table 1: Literature review

author(s)	year	Data source	Method	Results	Comments
Zakarya Farou, Noureddine Mouhoub and Tomas Horvath	2020	DNA microarray data: Colon cancer dataset and Breast cancer dataset.	Used method is data augmentation technique to study the effect of Gene Expression Generator (GEG) on classification accuracy. The classifiers used are Support Vector Machine (SVM), K-nearest Neighbors (KNN) and Decision Trees (DT).	GEG improved the classification accuracy compared to the baseline (original data only).	Based on the good results obtained, GEG can be considered as a promising data augmentation technique to improve the classification accuracy.
Andrew Beers, James Brown, Ken Chang, J. Peter Campbell, Susan Ostmo, Michael F. Chiang and Jayashree Kalpathy-Cramer	2018	Retinopathy of Prematurity Dataset and Brain Tumor Magnetic Resonance Imaging Dataset	Progressively Growing GAN (PGGAN) method. The GAN is trained to produce realistic down sampled images at 4x4 pixel resolution from a 128-vector latent space via a pair of convolutional layers, and then discriminate between real and synthesize images via a symmetric pair of convolutional layers in the discriminator.	Image Quality: high quality images are produced In both retinal fundus images and glioma multi-modality MRI images. Effect of Segmentation Channels: images trained without segmentation maps often obscured details important to pathological diagnosis.	The latent space of GANs often encodes semantic information about the images produced, and that latent vectors similar to each other in latent space produce qualitatively similar output images. That could be useful for generating images of a certain phenotype.

Hoo-ChangShin,Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine Andriole, and Mark Michalski	2018	Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)dataset	For brain segmentation, the generator G is given a T1-weighted image of ADNI as input and is trained to produce a brain mask with white matter, grey matter and CSF. The discriminator D on the otherhand, is trained to distinguish "real" labels versus synthetically generated "fake" labels.	A much improved performance with the addition of synthetic data is observed without usual data augmentation (GAN-based (no aug)). However, a small increase in performance is observed when added with usual data augmentation.	They believe that the generated synthetic images having half the resolution, coupled with the lack of the image sequences for training other than T1-weighted ones possibly led to the relatively small increase in segmentation performance compared to using the usual data augmentation techniques.
Poonam Chaudhari, Himanshu Agrawal and Ketan Kotecha	2019	Gene expression microarray data from datasets publicly available on the NCBI repository (https://www.ncbi.nlm.nih.gov/).	They used three main methods: KNN, Basic GAN and Modified generator GAN (MG-GAN). MG-GAN: This is the proposed approach where the generator is also trained with the help of original data, to generate synthetic samples adjacent to original samples.	The loss value decreases from 0.6978 at the start to 0.0082 by the end of 20,000 epochs. Thus, the augmented data generated through this approach are extremely close to the original data.	MG-GAN showed improvements in accuracy more than KNN or basic GAN. MG-GAN has a better recall value for all the datasets. MG-GAN approach is suitable for critical and sensitive application areas like medical data.

Talha Iqbal, Hazrat ALi	2018	DRIVE dataset and STARE dataset. These both datasets include a broad spectrum of vascular structured retinal images.	They used MI-GAN, Commonly used technique of encoder-decoder is adopted here. For segmentation, they utilize gold standard segmented images. They used Recent advancement in image style transfer, the synthesized image is based on a particular style representation provided by input.	They generate segmented images using ground truth segmented images of each dataset. The proposed method generates concordant probability values to the gold standard.	The result is enough to claim that a powerful discriminator y framework is key for successful training of the networks with GANs. Their method outperformed all the existing methods and shows better dice coefficient and AUC values.
Ramon Vinas, Helena Andres, Terre PietroLio, Kevin Bryson	2021	E.coli microarray data and Human RNA-seq data across a broad range of tissue- and cancer types.	The method builds on a WGAN-GP. The generator aims at producing samples. The critic takes gene expression samples from two input streams (the generator and the data distribution) and attempts to distinguish the true input source. They train the generator and the critic to solve a minimax game based on the Wasserstein distance.	The generated data from their model achieves the best realism scores, outperforming SynTReN and GNW by a large margin. They trained our model on a dataset that combines RNA-seq data from the GTEx and TCGA projects. Their model seems to capture tissue- and cancer-specific properties of transcriptomics data.	The GAN leverages real expression data to build a generative model in an unsupervised manner and does not require any information on the regulatory interactions.

Chapter 3 : Materials and methods

3.1. Generating Gene expression data:

3.1.1. Introduction to the generative adversarial model:

Our method is built on WGAN-GP proposed by Arjovsky et al. [4], which is an alternative to the traditional GAN. GANs, which are proposed by Goodfellow et al. [5], has quickly received great interest, as they are a powerful class of generative models that cast generative modeling as a game between two networks: a generator network produces synthetic data given some noise source and a discriminator network discriminates between the generator's output and true data, then it sends its feedback to the generator network which uses this feedback to improve the quality of the generated data, to bring them closer to the realistic data. The process goes on until the discriminator cannot differentiate between the real data and the original data.

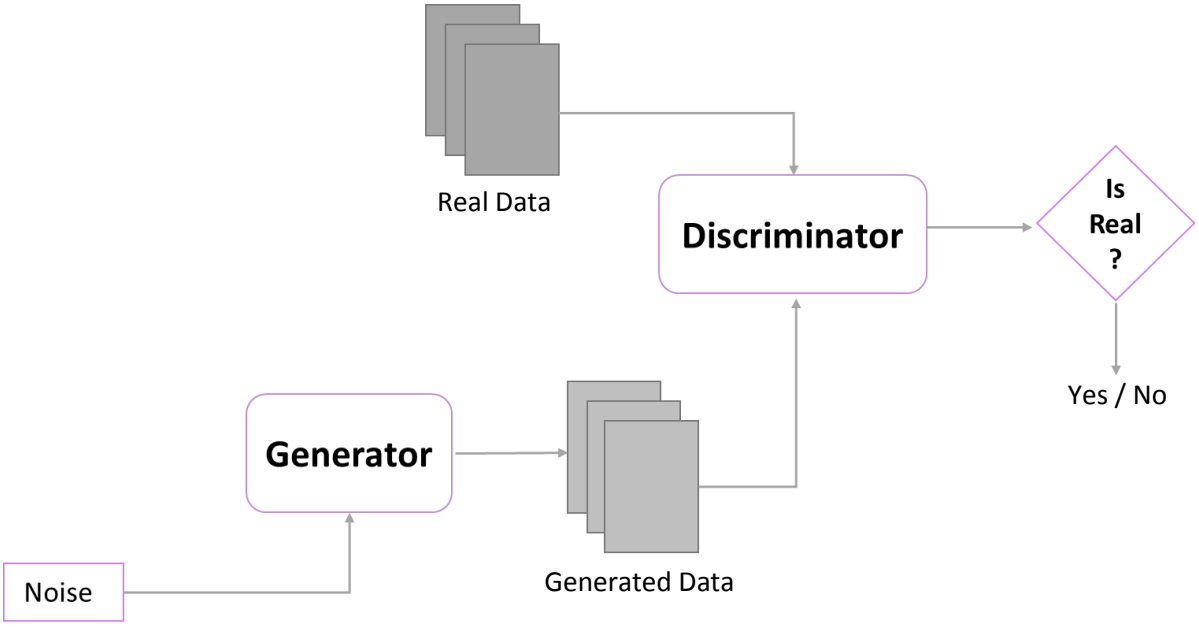

Figure 2: Representation of the generator and discriminator

Then WGAN was introduced by Arjovsky et al. [7], and it is similar to the traditional GAN models, as the generator receives input data with gaussian distribution *Z* and outputs data with distribution *Pg* and the discriminator receives them and the real data with distribution *Pr* and is trained to be able to differentiate between them. The main goal of the discriminator is to maximize the difference between the two distributions, while the main role of the generator is to approximate the distribution *Pg* to the distribution *Pr* [7]. The distance between *Pg* and *Pr* can be defined by many statistical ways, but in WGAN it is defined by Wasserstein distance, which is calculated by:

$$W(P_r, P_9) = E_{x\sim P_r}[D(x)] - E_{z\sim p(z)}[D(G(z))]$$

where *D* is the discriminator function, and it is called here the critic which is very similar to the discriminator but it doesn't have the sigmoid function at the end of the network and outputs a scalar score rather than limiting the output to be between 0 and 1 [7].

The generator and the critic are trained to solve the minmax game based on the Wasserstein distance, as the generator works on minimizing it and the critic's goal is to maximize it.

$$\min_G \max_D [\, E_{x\sim P_r}[D(x)] - E_{z\sim p(z)}[D(\, g_\Theta(z))] \,]$$

For providing smoother gradient and avoiding the vanishing gradient problem, the critic function *D* has to be 1-Lipschitz function, i.e. the norm of the critic's gradient with respect to x must be at most one everywhere [7,8]. To enforce the constraint, WGAN applies a very simple clipping to restrict the weights of the critic to be within a certain range controlled by a hyperparameters called *c*.

WGAN made progress toward stable training of GANs, but sometimes can still have some problems due to the use of wight clipping, So, Arjovsky et al. [4] proposed an alternative way for enforcing the constraint, which is penalizing the model if the gradient norm moves away from its target norm value 1.

By updating the previous minmax game with the gradient penalty we get:

- Generator's loss function:

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

- Critic's loss function:

$$\underbrace{E_{\hat{x}\sim P_g}[D(\hat{x})] - E_{x\sim P_r}[D(x)]}_{\text{Original critic loss}} \; \underbrace{\lambda E_{\bar{x}\sim P}[||\nabla_{\bar{x}} D(\bar{x})||_2 - 1]^2}_{\text{The gradient penalty}}$$

Where λ is a user-definable hyperparameter, and $\bar{x}$ is a random point along the straight line that connects x and $\hat{x}$ [8].

3.1.2. Datasets:

3.1.2.1. 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 [2]. We downloaded mRNAseq_693 dataset [13,14] 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 [3]. 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.

3.1.2.2. KIRC dataset:

Kidney renal clear cell carcinoma (KIRC) is one of the most common cancers with high mortality all over the world. We worked with the TCGA-KIRC data set for kidney cancer from The Cancer Genome Atlas (TCGA)I, t 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.

3.1.3. Data preprocessing:

3.1.3.1. Gene selection:

First, we need to select the genes we are interested in, we have two ways for gene selection, which are Getting the differentially expressed genes (DEGs) and detecting the disease pathway genes.

3.1.3.1.1. Getting the differentially expressed genes (DEGs):

Differentially expressed genes (DEGs) are the genes whose gene expression levels differ from one condition to another, i.e. a gene is declared as differentially expressed if an observed difference in the expression levels between two conditions is statistically significant [10].

1. **Data Filtration:**

We removed the samples which have zero expression level in both the normal and cancerous data, then we removed genes that has zero values in all samples in the normal and cancerous data.

2. **Hypothesis testing:**

- We made a hypothesis test whose null hypothesis is that the distribution of the healthy samples is the same as the distribution of the diseased samples. This was a paired difference test, so we used the Wilcoxon signed rank test (the non-parametric version of the paired Student t-test) as we have paired samples and the pairwise difference between the diseased samples and the healthy samples is not normally distributed.
- Each gene will have a separate hypothesis test and so a separate p-value. We got all the p-values for all the genes
- We made a multiple hypothesis testing using the Bonferroni method. Bonferroni correction method redefines the significance level to be $\frac{\alpha}{m}$, where m is the number of comparisons.
- We compared the corrected P-values to the original significance level and kept the genes with corrected p-values <= original significance level. These genes are the DEGs using the hypothesis testing approach.

3. **Fold Change:**

Fold Change (FC) is a quantitative measure for the change in the expression levels under two different conditions. It computes a ratio score between the average diseased GE levels and the normal GE levels according to this equation:

$$FC = \frac{Average\ of\ the\ gene\ expression\ level\ under\ the\ normal\ condition}{Average\ of\ the\ gene\ expression\ level\ under\ the\ diseased\ condition}$$

We computed the average GE level of the healthy samples and the average GE level of the diseased samples. Then we calculated the FC and get the log-transform to the base2 of FC: log2(FC).

If |log2(FC)| ≥ log2(T), then the difference in the expression levels is significant, where T is a threshold for significance of the difference. Otherwise, the difference is not significant, and we can consider both levels to be of the same category.

4. **Intersection between the DEGs of hypothesis test and fold change method:** 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:**
   We visualized the DEGs using volcano plot. A volcano plot is a type of scatterplot that shows statistical significance (P value) versus magnitude of change (fold change). It enables quick visual identification of genes with large fold changes that are also statistically significant. These may be the most biologically significant genes.

3.1.3.1.2. Getting the disease pathway genes:

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 [9].

3.1.3.2. Data normalization:

To make the learning rate easier, we standardize the expression values to keep them within a reasonable range. If we let X to be an m*n matrix of expression values, where m is the number of samples and n is the number of genes [8]. We standardize the data as follows:

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

Where:

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

Where $\boldsymbol{Z_{i,j}}$, $\boldsymbol{\mu_j}$, $\boldsymbol{\sigma_j}$ represents the standardized expression, the mean and the standard deviation of each gene respectively. The expression values of each gene in the resulting standardized expression matrix Z have zero mean and unit standard deviation.

3.1.4. The model architecture:

For our model, it is based on the WGAN-GP with 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:

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

- Critic's loss function:

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

3.1.5. Evaluating generated gene expression data:

It is a difficult task to assess the realism of a synthetic gene expression dataset, as the gene expression data are sensitive data. We used the quality assessment measures proposed by Ramon et al. [8], to evaluate our synthetic data. We calculated a similarity coefficient using Pearson's correlation coefficient. Let ***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 ***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:

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

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

3.1.5.1. Correlation between real and generated data distance matrices:

Let X and Z be two matrices containing m_1 and m_2 n-dimensional observations, respectively, sampled from the real and synthetic distributions. And we calculate the distance matrix D^X and D^Z for each of them. The coefficient γ (D^X, D^Z) measures whether the pairwise distances between genes from the real data are correlated with those from the synthetic data, as illustrated in Figure (3).

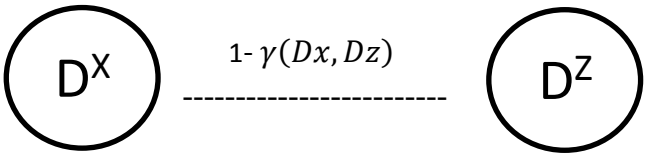

Figure 3: When the synthetic data resembles the real data, the distance matrices Dx and Dz are similar, and the distance 1- γ(D^X, D^Z) is close to 0.

3.1.5.2. Correlation between real and generated data distance dendrogram:

Dendrogram is a representation way for agglomerative hierarchical clustering according to a given linkage function, taking a n x n distance matrix as input and returning the n x n distance matrix of the resulting dendrogram. We define the real and generated dendrogrammatic distance matrices T^X and T^Z as:

$$T^X = C(D^X) \qquad T^Z = C(D^Z)$$

The coefficient $\gamma(T^X, T^Z)$ measures the structural similarity between the dendrograms, giving a score close to 1 when the real and generated dendrograms have a similar structure.

3.2. Generating medical images:

3.2.1. Generating abnormal brain tumor MRI images

Multi-parametric magnetic resonance images (MRIs) of abnormal brains (with tumor) are generated from segmentation masks of brain anatomy and tumor. This offers an automatable, low-cost source of diverse data that can be used to supplement the training set. For example, we can alter the tumor's size, change its location, or place a tumor in an otherwise healthy brain, to systematically have the image and the corresponding annotation. Furthermore, GAN trained on real medical images to generate synthetic images can be used to share the data outside of the institution, to be used as an anonymization tool. [11]

3.2.2. Generate glioblastoma brain MRI images:

Generating synthetic MRI brain images from segmented labels is considered to be an image-to-image translation problem. An image-to-image model, like pix2pix model described by Phillip et.,[12] can use the tumor segmented labels as inputs to train the generator, and the generator outputs synthetic abnormal brain image, which will be an input to the generator with a real abnormal brain image to discriminate whether the synthetic image is realistic or not.

Chapter 4 : Results

4.1. 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 2: 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

4.2. Analyzing generated gene expression data

4.2.1. Analyzing generated GBM gene expression data:

We made the evaluation based on the 12 most frequently mutant genes as described by Romana et al. [8]. We calculated the distance matrix for the real and the synthetic gene expressions for the 12 most frequent genes.

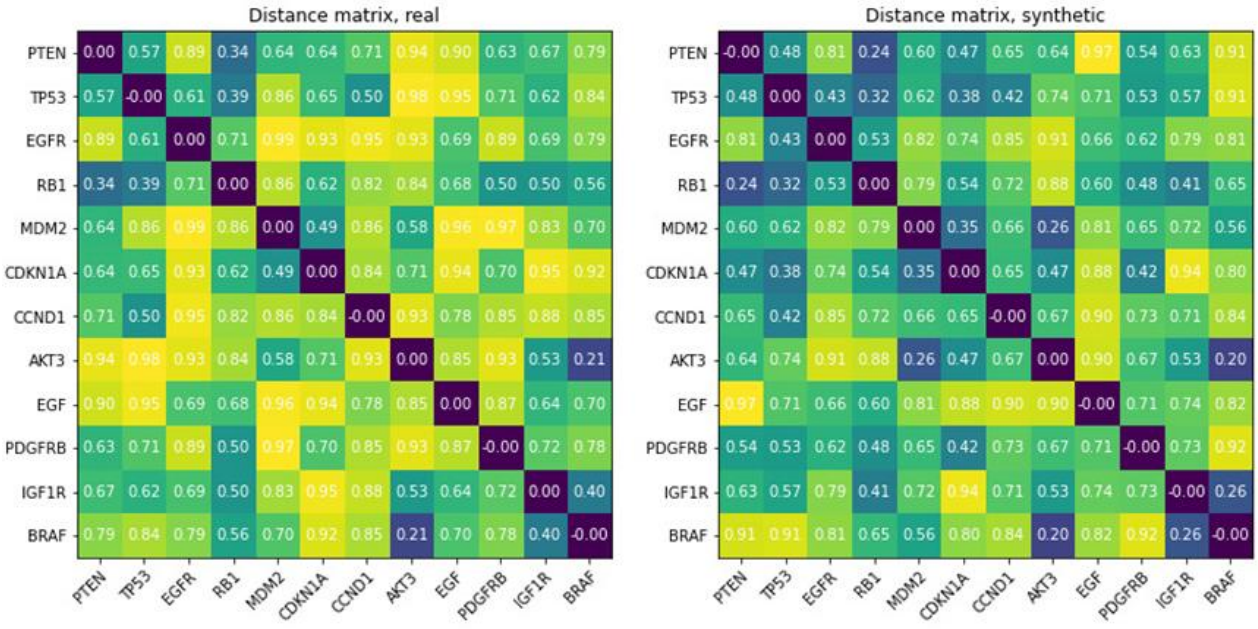

Figure 4: The Distance Matrix for real data on the left side and the distance matrix for generated data on the right side for the GBM carcinoma

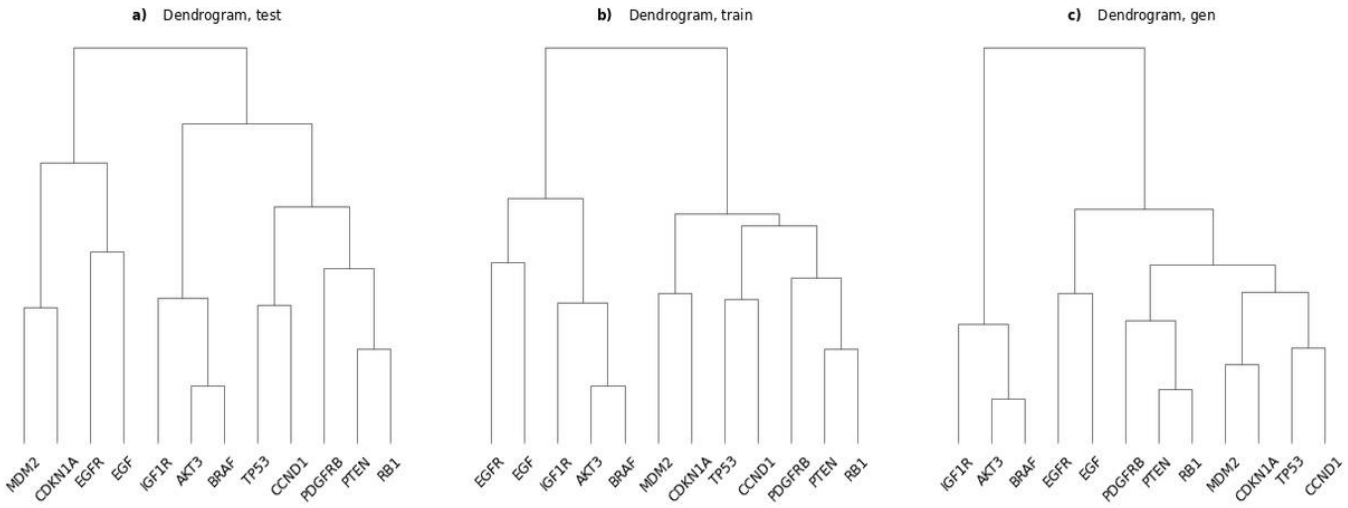

Figure 5: dendrograms for a subset of 12 GBM most frequently mutant genes.

The following figure shows the distribution of the real (black distribution) and generated GBM gene expression (blue distribution) for the 12 genes.

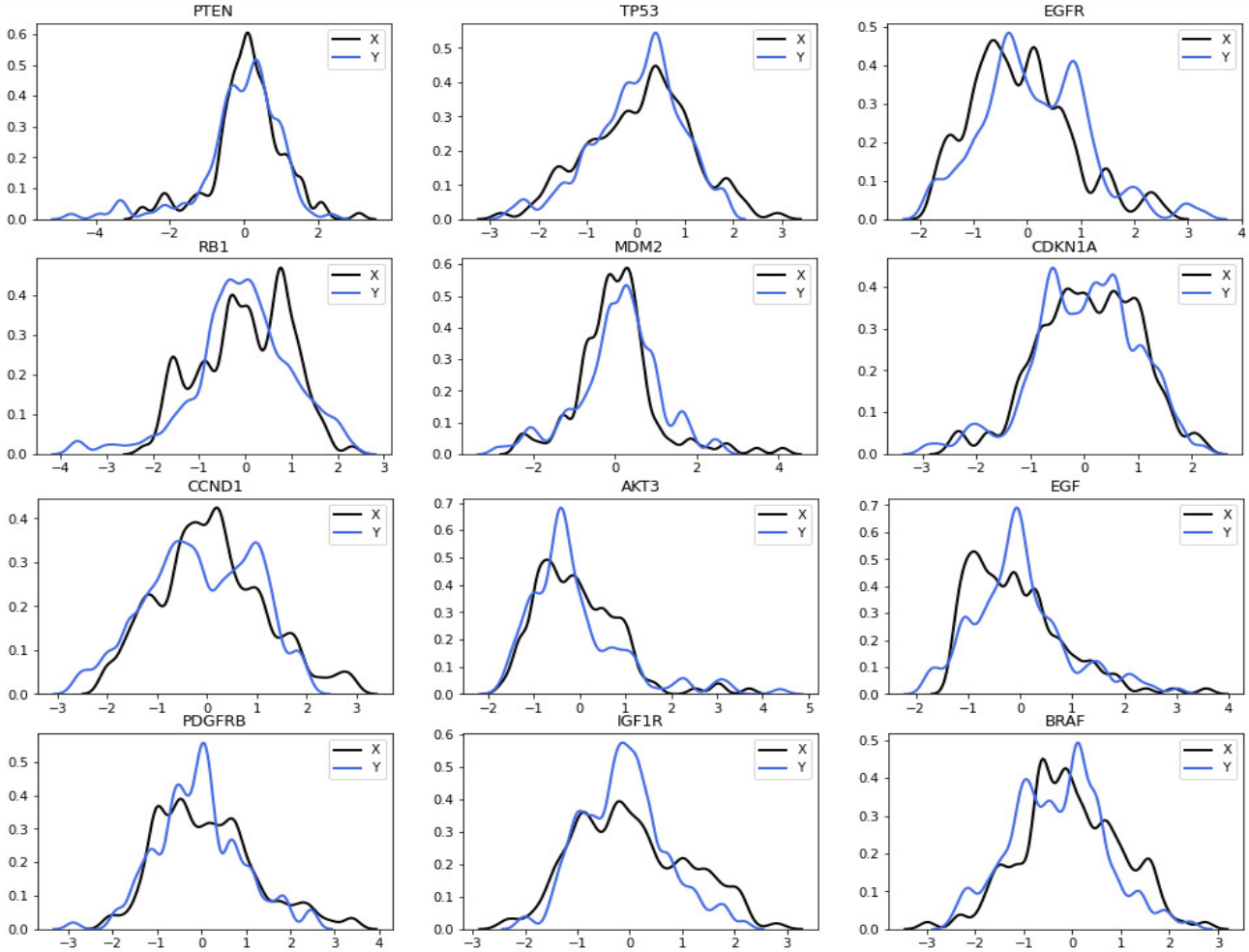

Figure 6 : The distribution of the real and generated GBM gene expressions

4.2.2. Visualizing the DEGs for KIRC data:

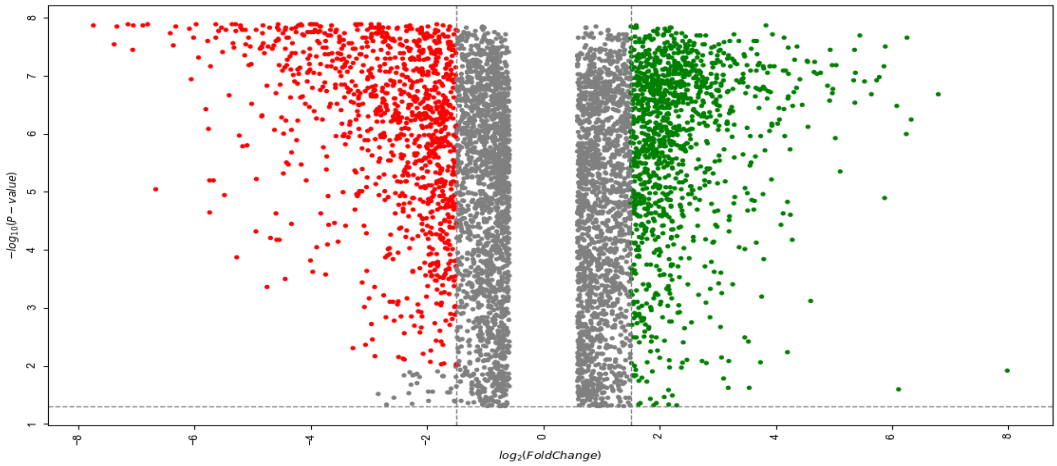

Figure 7: Volcano plot for visualizing DEGs of KIRC data

4.2.3. Analyzing generated KIRC gene expression data:

We visualize the analysis based on the first 12 KIRC genes in the disease pathway. We calculated the distance matrix for the real and the synthetic gene expressions for the 12 genes, as illustrated in the following figure:

Figure 8: the distance matrix for real data on the the left side and the distance matrix for the generated data on the right side for KIRC

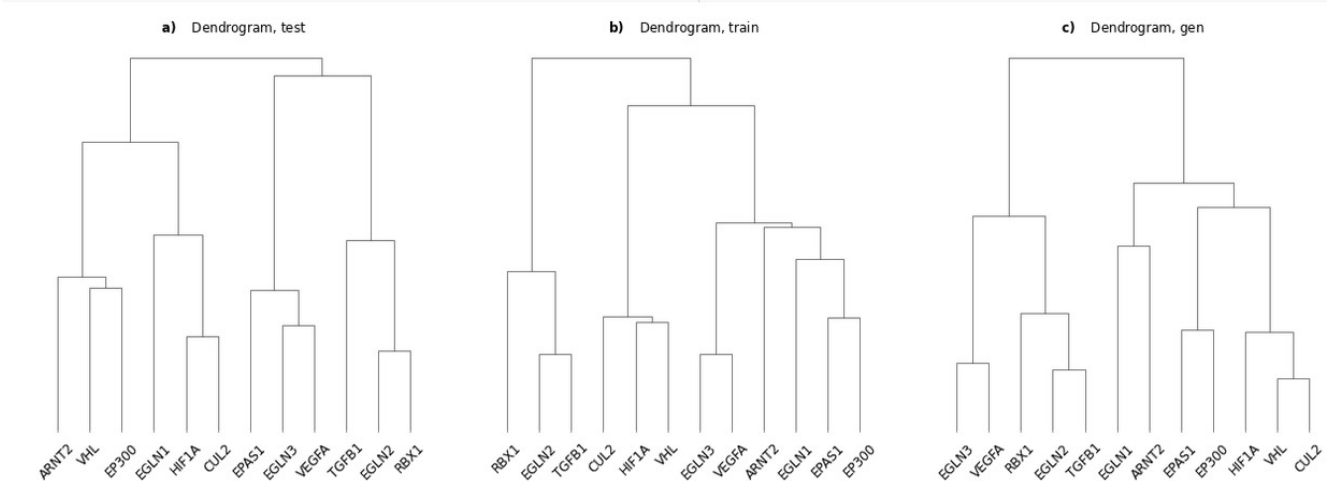

Figure 9: dendrograms for a subset of first 12 mutant genes in the KIRC disease pathways

The following figure shows the distribution of the real (black distribution) and generated KIRC gene expression (blue distribution) for the 12 genes.

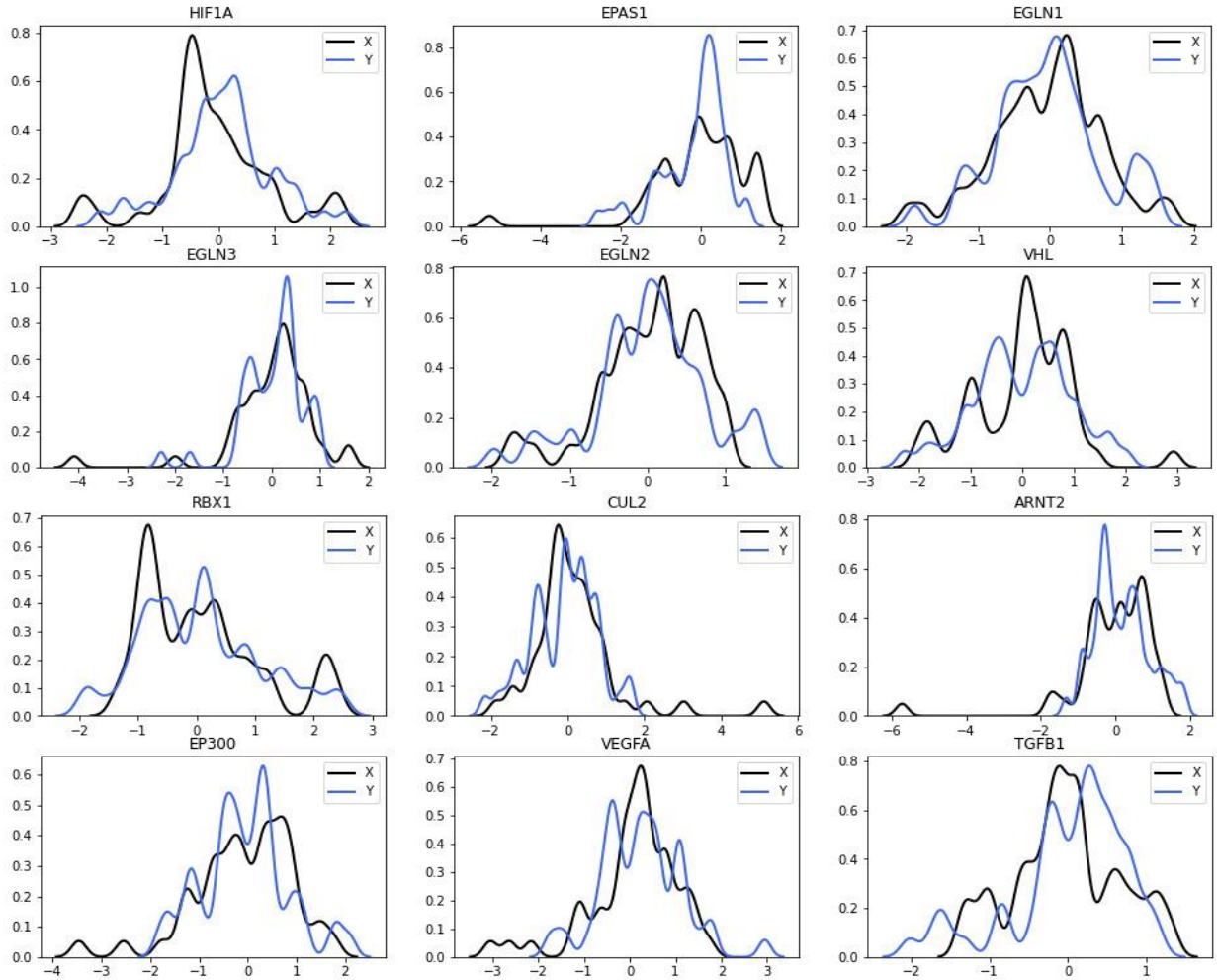

Figure 10: The distribution of the real and generated KIRC gene expressions

Chapter 5 Discussion

5.1. Gene expression data generation:

5.1.1. Datasets and preprocessing:

For the GBM datasets, we tried to use datasets from TCGA portal but the number of samples was limited. First, we downloaded the TCGA-GBM dataset (using the biolink library supported for R language) for the subjects having both genetic and imaging data in TCGA and TCIA respectively. But the total number of samples for the paired data was 72 sample, and for providing well training for the model we needed larger dataset. So, we used the dataset from the CGGA which has larger 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.

After getting the DEGs for KIRC dataset, we visualized the result by drawing the volcano plot between the log2(FC) on the x axis and -log10(P-Value) on the y axis as illustrated in Figure (7). The gray area shows the significant genes based on the hypothesis testing method and the non-significant genes based on the fold change method. The red area represents the area where the gene expression level decreased in the diseased data compared to the normal data. The green area represents the area where the gene expression level increased in the diseased data compared to the normal data. The red and green areas are the significant DEGs for the intersection between the fold change and hypothesis test.

5.1.2. The generative adversarial model:

We used the WGAN-GP for building our model, as the traditional GANs show problems with convergence, like Failure of the generator to improve due to the vanishing gradient problem, at which the gradient diminishes to the point that the generator cannot usefully learn from it, and mode collapse that occurs when the generator learns to only output specific classes from the distribution of the real data [6]. So, WGAN showed that the stability of learning can be improved and problems like mode collapse and vanishing gradients can be got rid of, as shown in Figure (11).

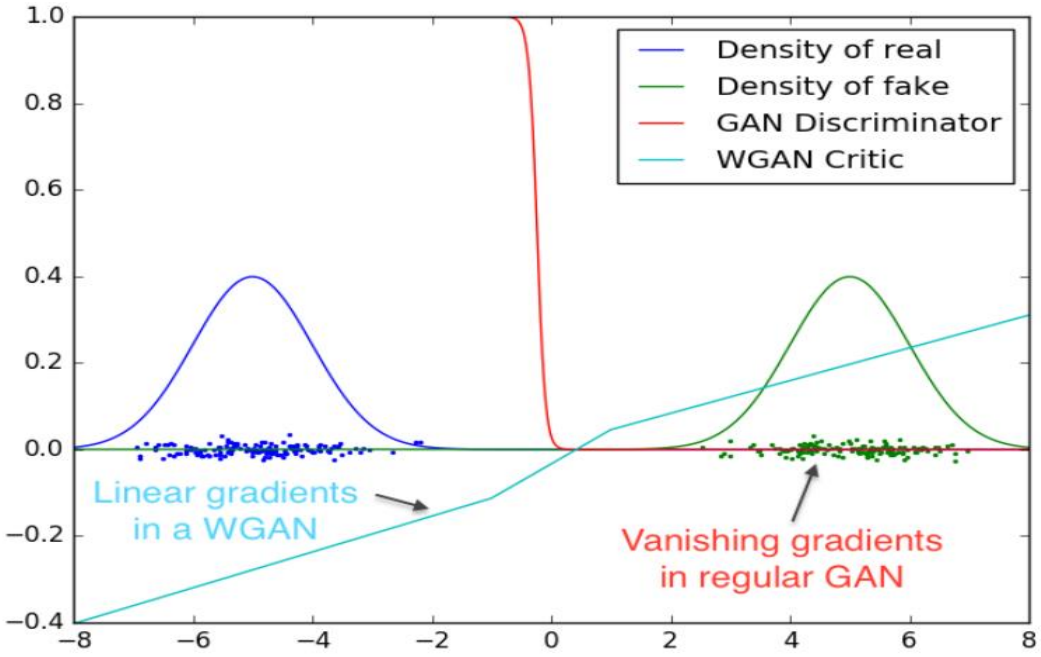

Figure 11: The discriminator in the traditional GAN results in vanishing gradients, while the critic in WGAN provides smooth gradient everywhere.

Despite providing more stability, WGAN can still have some problems, like generating poor samples, and these problems are often due to the use of weight clipping for enforcing the 1-Lipschitz constraint. Fortunately, WGAN-GP provided an alternative way for enforcing the constraint without using the weight clipping, which is penalizing the model if the gradient norm moves away from its target norm value 1. So, WGAN-GP was a good choice for building the model.

Checking the performance of the model is a necessary step for avoiding overfitting, as we can't know how well the performance of the model will be on new data until it is actually tested. To handle this point, 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, the dataset was divided into 5 equal parts, then:

1. A part was taken as test set
2. The remaining parts were taken as training set
3. The model was fit on the training set and evaluated on the test set
4. The evaluation score was retained and the model was discarded.

The previous steps were repeated 5 times with selecting the part next to the previously selected one in step (1) as the new test data. Then we got the retained scores to get the model score, which is the average of all the 5 retained scores and it was 92%.

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 as illustrated in Table (1). This is caused by the large number of genes selected from the DEGs compared to the number samples.

For analyzing the correlation between the real and generated data, we constructed distance matrices and dendrograms between the real and generated data as show in Figures (4) and (5) for GBM data, and Figures (8) and (9) for KIRC data. We deduced that the model closely matches the correlation and clustering expression patterns. And for illustrating the difference between the two distributions *Pg and Pr*, we visualized the *Pg and Pr* as illustrated in Figures (6) and Figure (10) for GBM and KIRC data respectively. This visualization represents small variations between the two distributions.

5.2. Medical Images Generation:

We searched a lot for a publicly available dataset for glioblastoma brain images that:

- has information about the tumor labels, and
- its subjects are the same as the subjects in the gene expression data, required for the mapping between the imaging and genetic data.

Our project meant to increase the available datasets provided for Radiogenomic studies by generating fake genetic and imaging data, but we couldn't reach a publicly available dataset with paired genetic and imaging data or brain tumor datasets that have tumor information, as till now the segmentation process is done manually by an expert radiologist.

Chapter 6 Conclusions and future work

6.1. Conclusion:

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 which depends on calculating the Wasserstein distance as it is the distance between the distribution of the generated data by the generator *Pg* and the distribution of the real data *Pr*. 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.

To evaluate the quality of the data generated by the model, we used visualization methods for detecting the difference between the generated data and the real data. We constructed the distance matrices and dendrograms for both the real and the generated data to get the correlation between them. Then we draw the two distributions *Pg and Pr* for each gene to understand the variations between the two distributions, and the variations were relatively small.

The visualized results indicates that the model closely matches the correlation and clustering expression patterns, and the distribution *Pg* is close to the distribution *Pr*, which illustrates that the model managed to generate realistic gene expression data.

6.2. Future work:

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.

Till now there is no solid knowledge about mapping between the phenotype and the genotype. But in the future, we may have large scale of researches in this topic, hence we will try to predict one datatype from the other.

References

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. 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.

3. 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.

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

5. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

6. Barnett SA. Convergence problems with generative adversarial networks (gans). arXiv preprint arXiv:1806.11382. 2018 Jun 29.

7. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. InInternational conference on machine learning 2017 Jul 17 (pp. 214-223). PMLR.

8. Viñas Torné R, Andrés-Terré H, Lio P, Bryson K. Adversarial generation of gene expression data.

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

10. 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.

11. Shin HC. et al. (2018) Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In: Gooya A., Goksel O., Oguz I., Burgos N. (eds) Simulation and Synthesis in Medical Imaging. SASHIMI 2018. Lecture Notes in Computer Science, vol 11037. Springer, Cham.

12. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

13. 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.

14. 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.