Bioinformatics

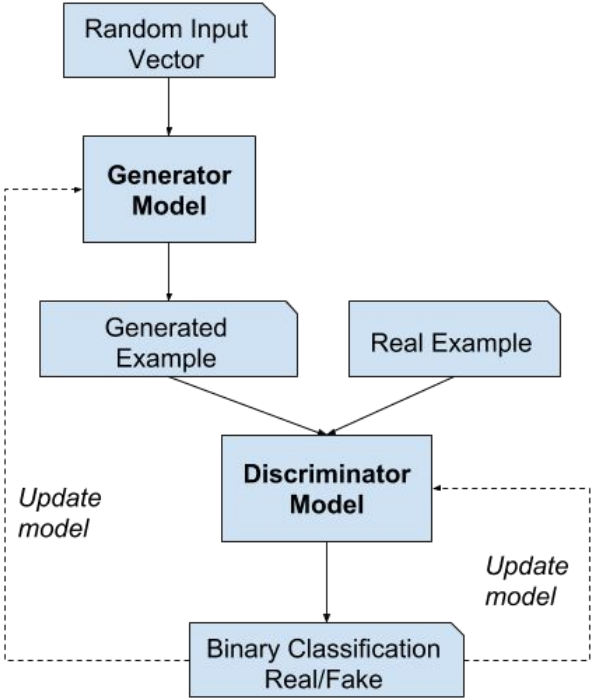
Subtype-GAN: cancer subtyping of multi-omics data

Project #3:
Tommaso Calò
Alessandro Desole
Valerio Di Eugenio
Roberto Franceschi

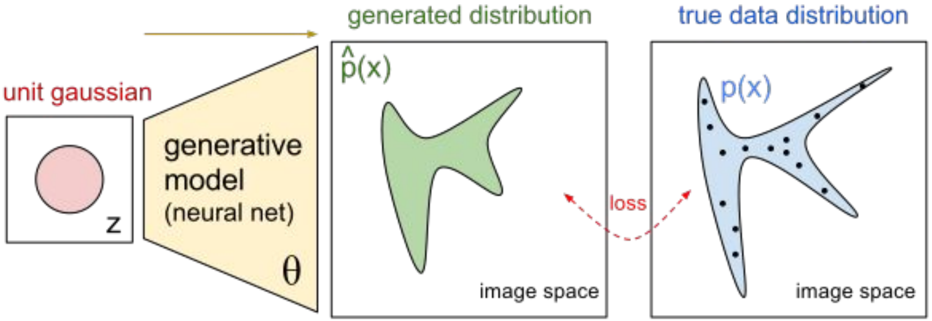
Outline

1. Introduction
    - GAN
    - Problem description
2. Dataset
3. Architecture
4. Results
5. Improvement images
6. Conclusions

Generative adversarial network (GAN)



Given a training set, GANs learns to generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic to human observers, having many realistic characteristics.



Vanilla GAN on synthetic dataset

- Synthetic dataset
  - 5000 samples equally divided in 5 clusters

- One GAN for each cluster of patient (cluster_id)

  NUM_EPOCHS = 500
  BATCH_SIZE = 32
  LEARNING RATE = 3e-4

- Evaluation of every GAN

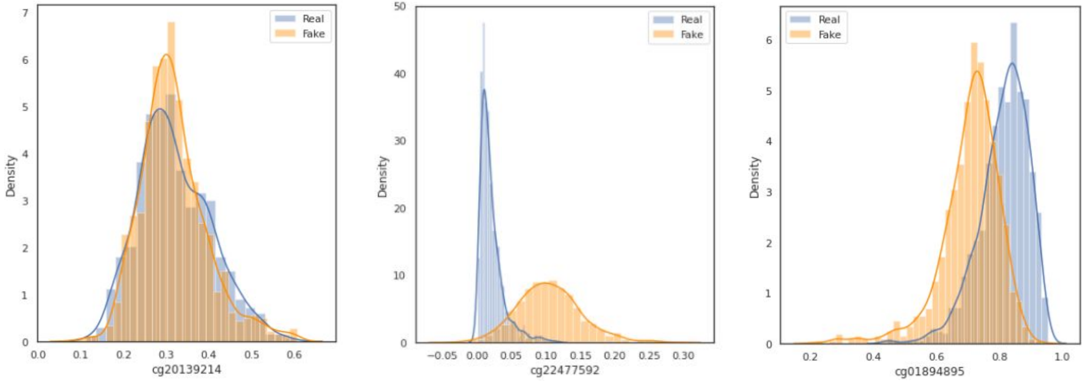
Limits Vanilla-GAN:

- Long training, sensible to hyperparameters
- Unstable results

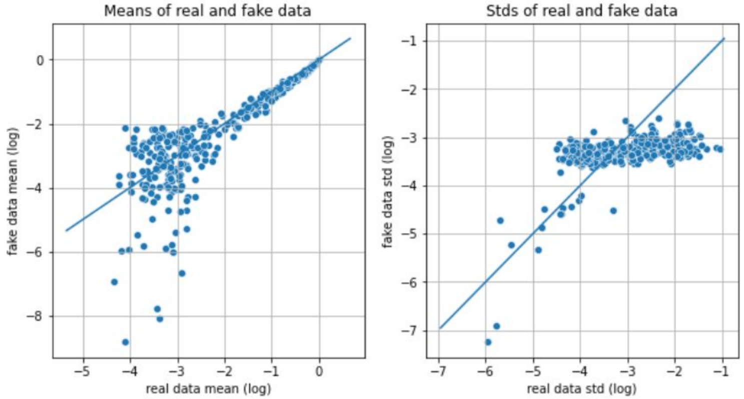
Results Vanilla GAN

- Results showed that Vanilla-GAN is not a feasible approach for the synthetic augmentation of the dataset
- *Kolmogorov–Smirnov test* used as quantitative indicator
  - H_0: two samples (real, fake) are drawn from the same distribution
  - null hypothesis rejected for most of the features

Distribution of some features (meth) for real and fake data. For some features the two distributions do not overlap enough highlighting different statistics (mean and std)



Complete view of the mean and std for meth data. Ideally mean and std of the features should lie on the bisector.



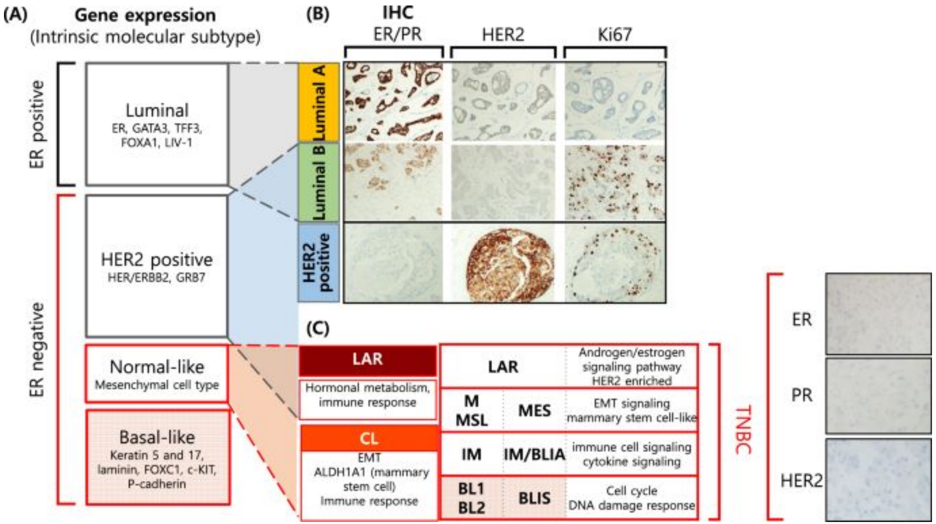
Problem description

Cancer Subtype

Describes the smaller groups that a type of cancer can be divided into, based on certain characteristics of the cancer cells. These characteristics include how the cancer cells look and specific gene expressions. It is important to know the subtype of a cancer in order to plan treatment and determine prognosis.

Objective

Due to the diversity and complexity of multi-omics data, it is challenging to develop integrated clustering algorithms for tumor molecular subtyping.
Our objective is to classify lungs cancer subtypes given multi-omics data exploiting the power of Generative adversarial networks.

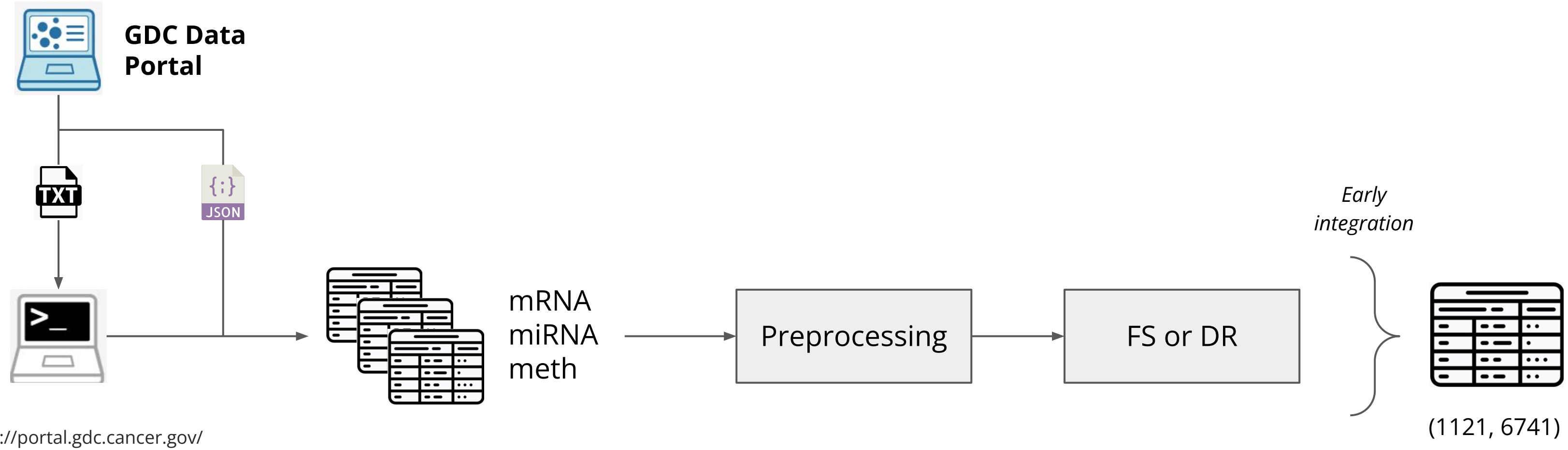


Example of cancer subtypes classification (breast cancer)

Dataset

- Multi-omics: *mRNA, miRNA, meth*
- Data downloaded from GDC portal
- Preprocessing (NA, log2, scaler)
  - Feature selection
  - Dimensionality reduction
- *Early integration approach*

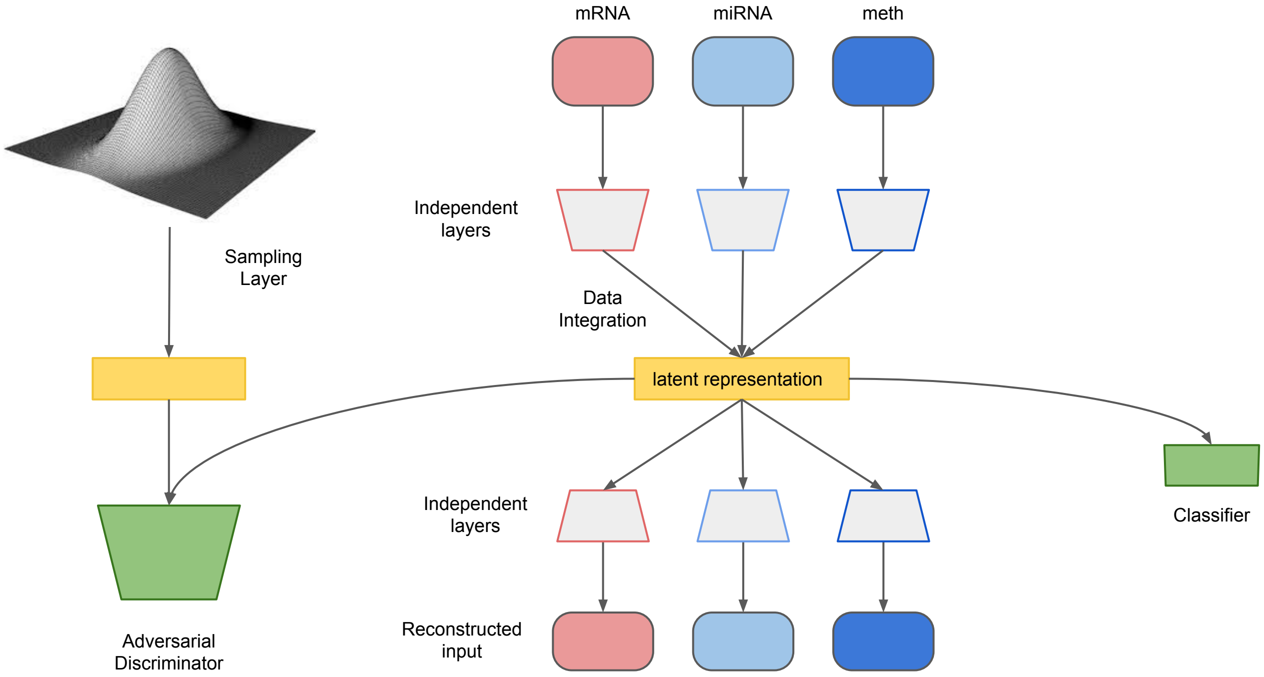
- Number of samples: 1121
- Subtypes:
  - TCGA-LUSC
  - TCGA-LUAD
- Classes: 4 (Subtype + Healthy/Tumor)



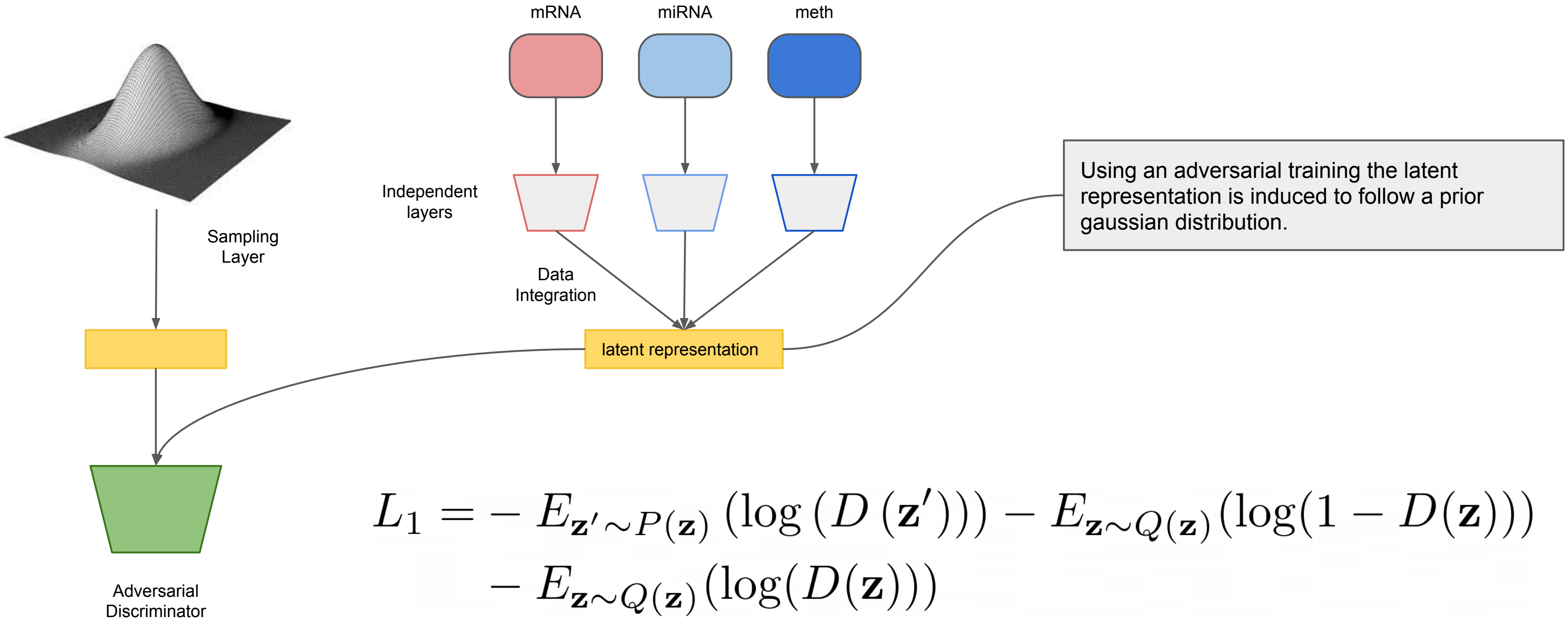
GDC Dataset, https://portal.gdc.cancer.gov/

Network architecture - Objective

The aim is to classify latent representation derived by exploiting relevant features of multi-omics data



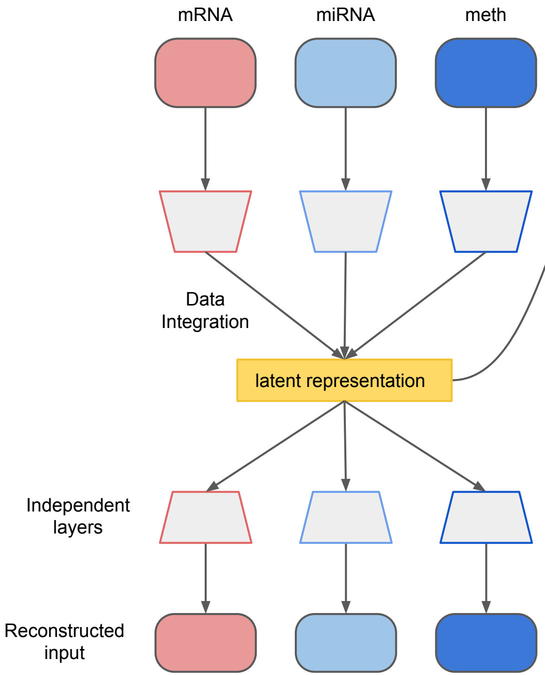
Network architecture - Adversarial training



Using an adversarial training the latent representation is induced to follow a prior gaussian distribution.

$$L_1 = -E_{\mathbf{z}' \sim P(\mathbf{z})}\left(\log\left(D\left(\mathbf{z}'\right)\right)\right) - E_{\mathbf{z} \sim Q(\mathbf{z})}(\log(1 - D(\mathbf{z}))) \\ - E_{\mathbf{z} \sim Q(\mathbf{z})}(\log(D(\mathbf{z})))$$

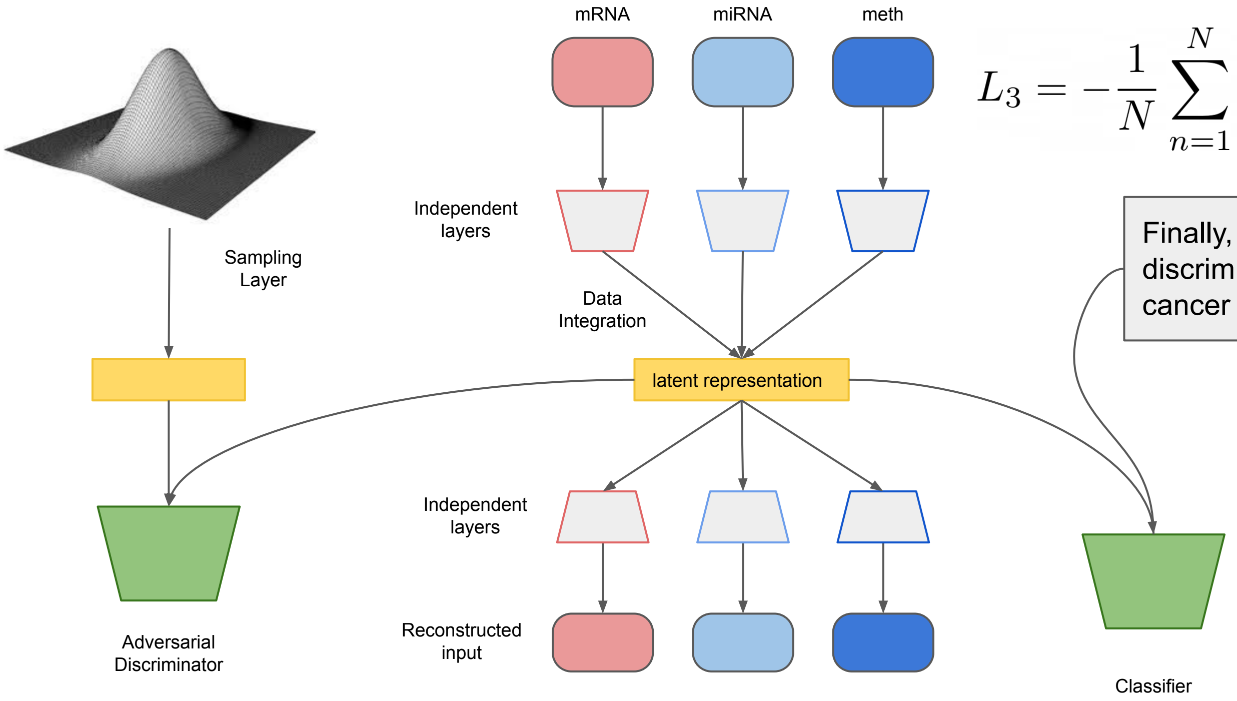
Network architecture - Reconstruction



In order for the latent representation to keep relevant information about the original distribution a reconstruction loss is introduced.

$$L_2 = \frac{1}{mn} \sum_{k=1}^{m} \sum_{i=1}^{n} \|\mathbf{x}_{ki} - \mathbf{x\prime}_{ki}\|_2^2$$

Network architecture - Classifier



$$L_3 = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right]$$

Finally, a classifier discriminates between the cancer subtypes

Loss Functions

$$L_1 = -\,E_{\mathbf{z}' \sim P(\mathbf{z})}\left(\log\left(D\left(\mathbf{z}'\right)\right)\right) - E_{\mathbf{z} \sim Q(\mathbf{z})}(\log(1 - D(\mathbf{z}))) \\ -\,E_{\mathbf{z} \sim Q(\mathbf{z})}(\log(D(\mathbf{z})))$$

Adversarial training Loss (BCE)

$$L_2 = \frac{1}{mn} \sum_{k=1}^{m} \sum_{i=1}^{n} \left\| \mathbf{x}_{ki} - \mathbf{x}'_{ki} \right\|_2^2$$

Reconstruction loss (MSE)

$$L_3 = -\frac{1}{N} \sum_{n=1}^{N} \left[y_n \log \hat{y}_n + (1 - y_n) \log \left(1 - \hat{y}_n\right) \right]$$

Classification loss (CLE)

$$L = L_1 + L_2 + L_3$$

Total loss

Hyperparameter tuning

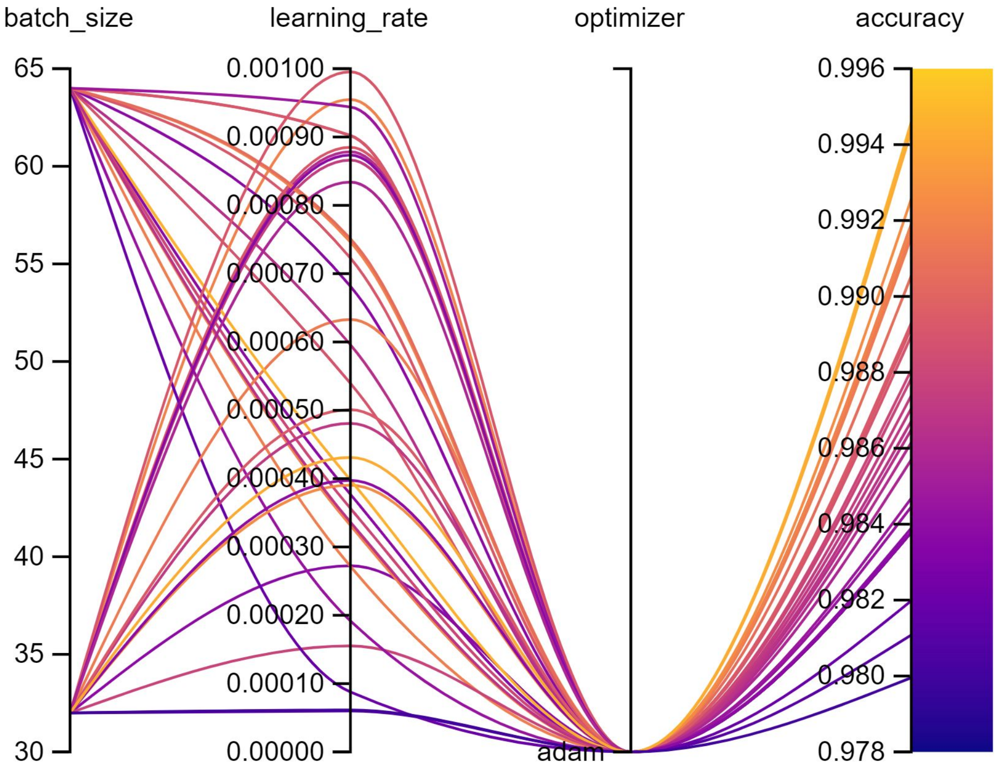
Tuned hyperparameters:

- Learning rate
- Batch size
- Optimizer
- Epochs

Best configuration found:

- Learning rate = 0.00043
- Batch size = 32
- Optimizer = 'adam'
- Epochs = 100

▶ The graph reports the top-30 runs according to the accuracy on the test set.



Ablation Study

Different preprocessing:

- *K best*
- *PCA*
- *Ensemble method for feature selection*
    - 3 classifiers with major voting

Different backbone:

- *Classifier*
- *Classifier + Encoder*

Results overview

Results computed considering 2 subtypes (LUAD and LUSC), i.e., 4 classes:

- Comparison among different preprocessing (feature selection methods)
- Comparison among different methods

	Model	Classifier + Encoder	Only Classifier
K Best	0.9811	0.9676	0.9495
PCA	0.9785	0.8919	0.8839
Ensemble Method	**0.9919**	0.9703	0.9521

The table reports the classification accuracy on the test dataset. The best model is the Subtype-GAN plus the classification layer. Moreover, the best feature selection method overall is the ensemble method.

Results - Extended dataset

Considered omnics: miRNA and mRNA

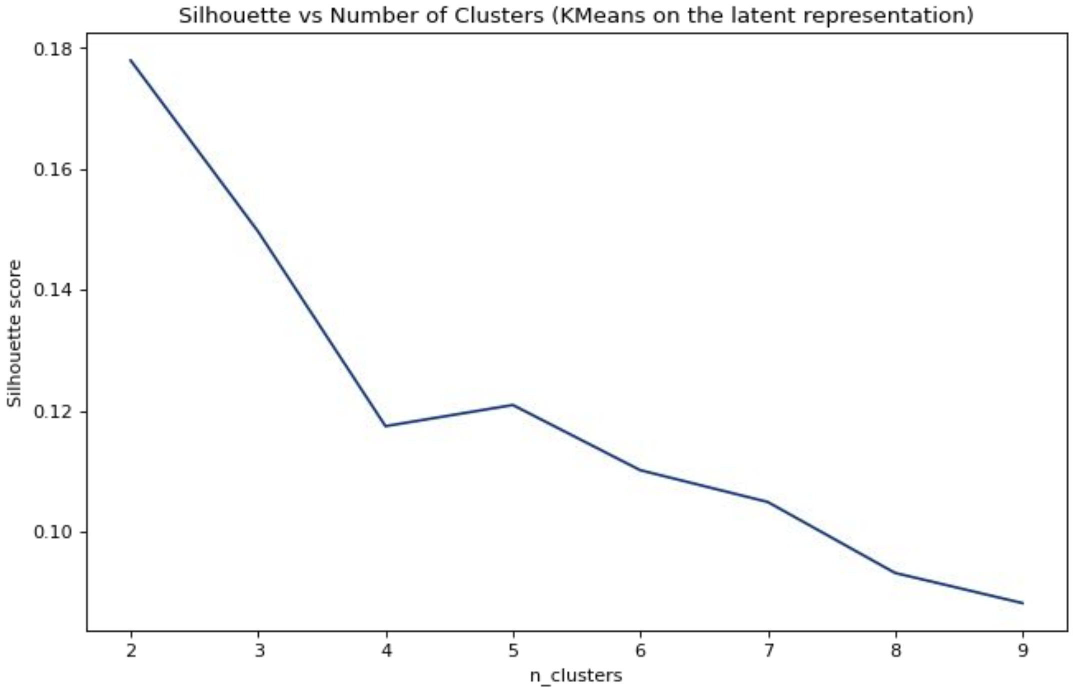
Results computed considering 5 subtypes, meaning 10 classes:

- TCGA-LUAD
- TCGA-LUSC
- CPTAC-3
- TCGA-BRCA
- TCGA-KIRC

	Model	**Classifier + Encoder**	**Only Classifier**
K Best	0.9792	0.9594	0.9422
PCA	0.9744	0.8887	0.8758
Ensemble Method	**0.9805**	0.9615	0.9391

The table reports the classification accuracy on the test dataset. Again the best model is the Subtype-GAN plus the classification layer.

Clustering on latent space



- We also tried to perform **clustering** directly on the latent representation in order to try a completely unsupervised approach.

- The results shows that latent representation is not easily clusterable since silhouette scores don't exhibit an elbow point.

- Rand index: 0.54

Improvement images

Unsupervised segmentation → segment an image into an arbitrary number of plausible regions without any previous knowledge.

Clustering on pixels according to three criteria:

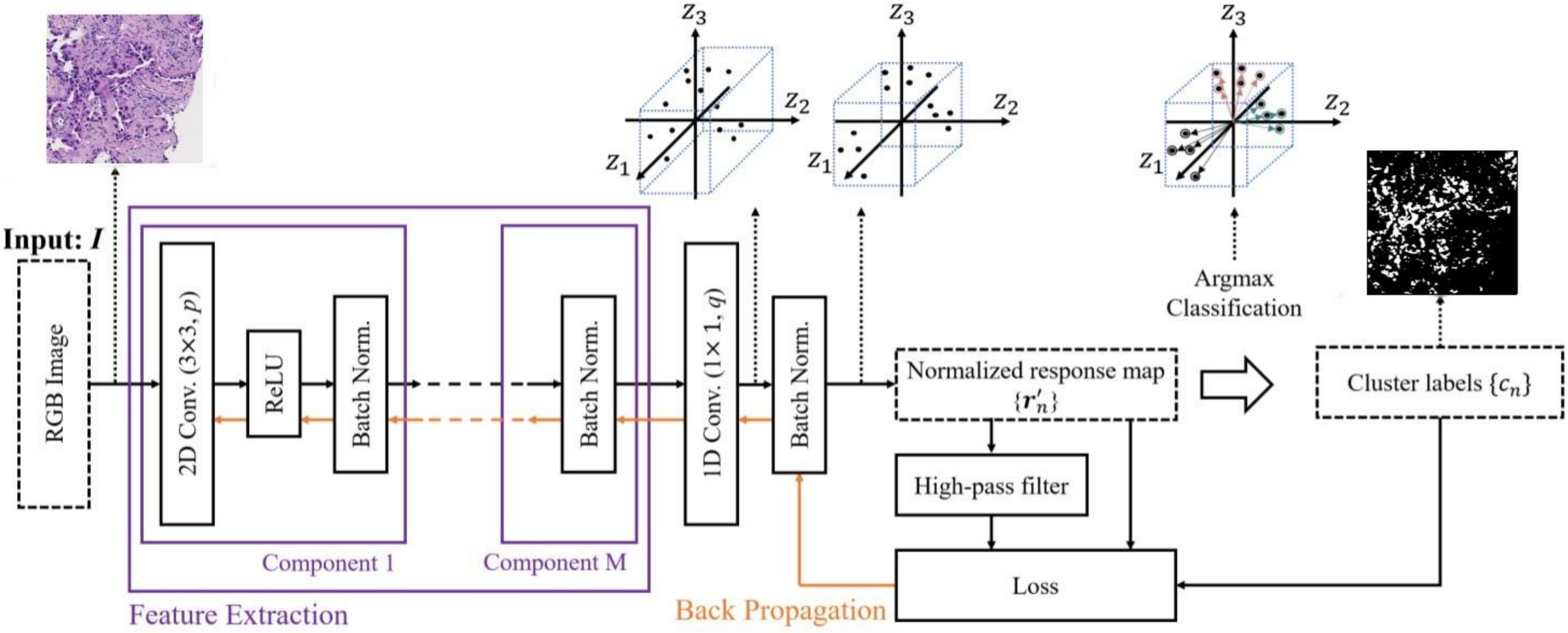
1. Pixels of similar features assigned to the same label
2. Spatially continuous pixels assigned to the same label
3. Large number of unique cluster labels

Method used:

→ CNN-based algorithm that jointly optimizes features extraction functions and clustering functions to satisfy these criteria.

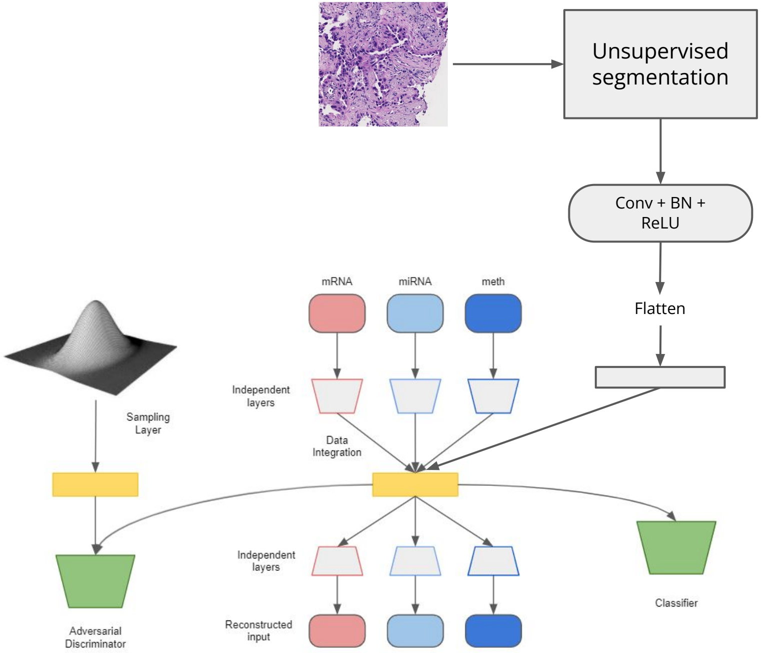
Kim, Wonjik et al. **Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering.** *IEEE Transactions on Image Processing* (2020).

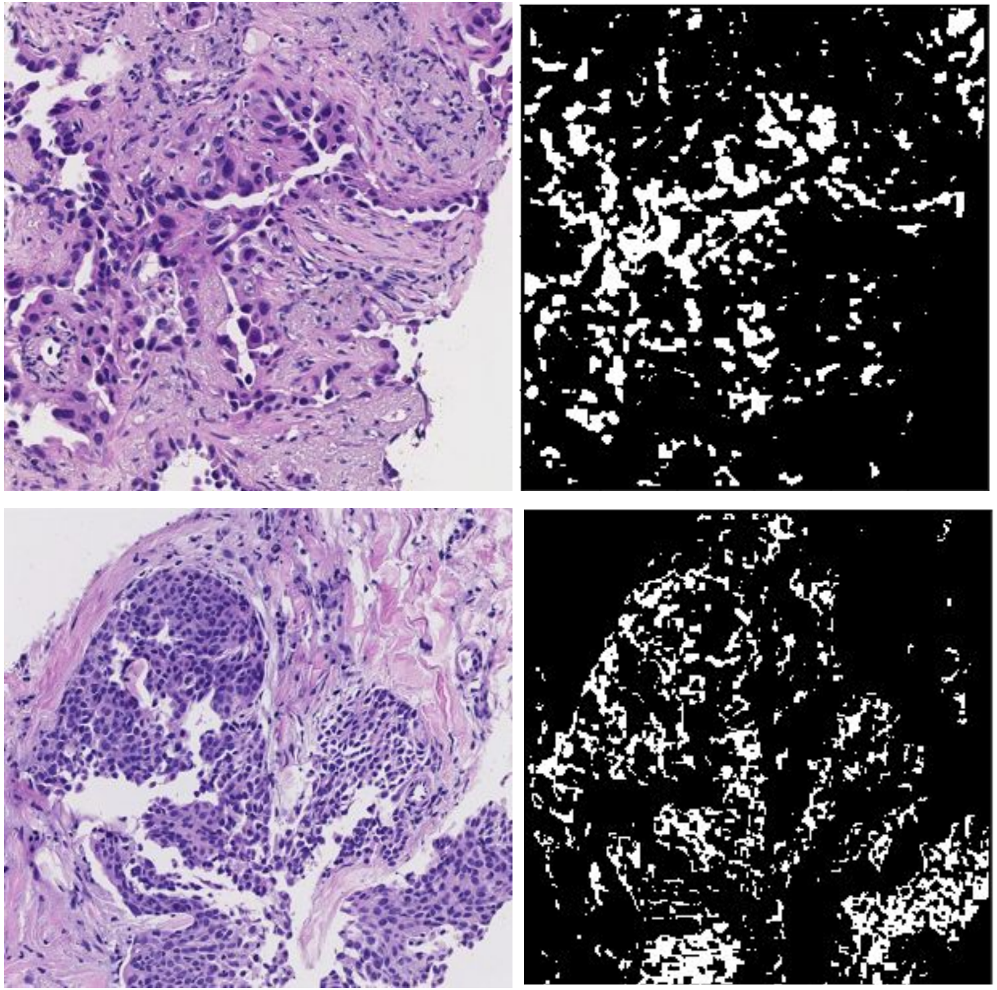
Improvement images



Kim, Wonjik et al. **Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering.** *IEEE Transactions on Image Processing* (2020).

Plug images





▶ Example of unsupervised segmentation on pathological images of adenocarcinoma (top) and squamous cell carcinoma (bottom) are displayed.

Conclusion

- Even if the accuracy is high we must consider that we are considering only two subtype (LUSC and LUAD in their healthy/tumor variants) meaning only 4 classes.

- Subtype-GAN is able to generalize better respect to the other method tested, regardless on the feature selection applied on the data.

- The preprocessing method have an influence on the final results. Highlighting the importance of extracting the most significant features from every omic.

- A possible future improvement is to add images as input of the model to try to improve the level of discrimination among subtypes.

Thank you for your attention!

Tommaso Calò
Alessandro Desole
Valerio Di Eugenio
Roberto Franceschi

Resources

1. Hai Yang, Rui Chen, Dongdong Li, Zhe Wang, *Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data*, Bioinformatics, Volume 37, Issue 16, 15 August 2021, Pages 2231–2237

2. Kim, Wonjik et al. *"Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering."* IEEE Transactions on Image Processing 29 (2020): 8055-8068.

3. Wang, S., Chen, A., Yang, L. et al. *Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome*. Sci Rep 8, 10393 (2018).

4. GDC Dataset: https://portal.gdc.cancer.gov/