Machine learning in translational oncology research

Franciszek Górski
Gradient Science Club, Multimedia Systems Department

About me

- Deep learning researcher from ETI Faculty of PG
- Interests: deep and machine learning for various problems including biometric verification, cancer classification, object detection
- Working in Multimedia Systems Department
- Collaborate with Center for Biostatistics and Bioinformatics from Gumed
- Chairman of Gradient Science Club

About the project

- Project with collaboration of Center for Biostatistics and Bioinformatics of Medical University of Gdansk

- Classification of so called cancer data

- Investigate the ability of various machine learning algorithms like boosted trees or neural networks

Our papers

- Sebastian Cygert, Franciszek Górski, Piotr Juszczyk, Sebastian Lewalski, Krzysztof Pastuszak, Andrzej Czyżewski, and Anna Supernat: *Towards Cancer Patients Classification Using Liquid Biopsy*
- Sebastian Cygert, Krzysztof Pastuszak, Franciszek Górski, Michał Sieczczynski, Piotr Juszczyk, Antoni Rutkowski, Sebastian Lewalski, Robert Rózanski, Maksym Jopek, Jacek Jassem, Andrzej Czyzewski, Thomas Würdinger, Myron G. Best, Anna J. Zaczek And Anna Supernat: *Platelet-based liquid biopsies through the lens of machine learning*

Tumor-educated Platelets

- A cancer in the patient's body is editing platelets
- It inserts information about itself into the platelets
- Platelets circulates around the organism
- Physicians can take blood sample and analyse it for the presence of cancer

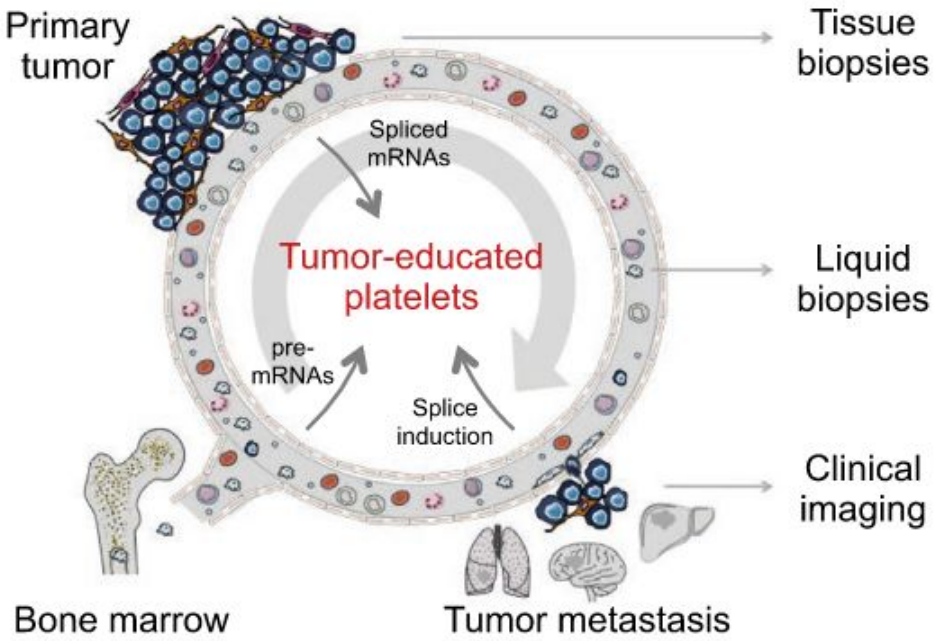


Figure from: Best et al. (2015). RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics

Liquid biopsy

- Minimally invasive method of gathering samples for cancer detection
- Getting interests thanks to Tumor-educated Platelets mechanism
- Much faster than tissue biopsy
- Could be easily preprocessed

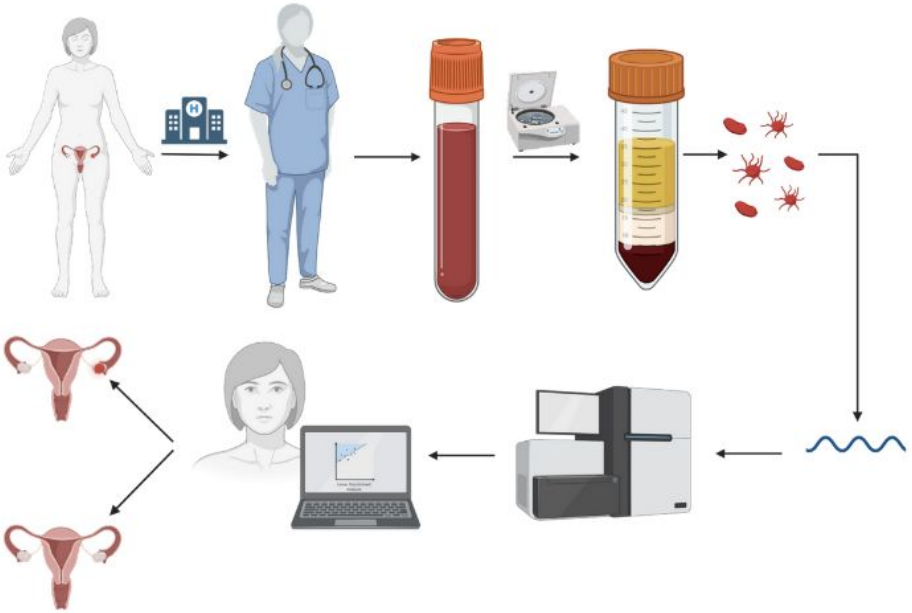


Figure from: Cygerts, et al. Platelet-based liquid biopsies through the lens of machine learning

Our approach

Our approach

We decided to implement two approaches to the classification of our data:

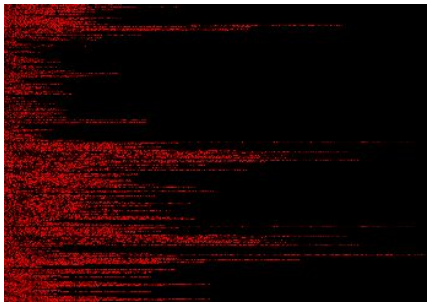
- Classify the data in the binary manner - class 0 means sample of patient without cancer, class 1 sample of patient with cancer.

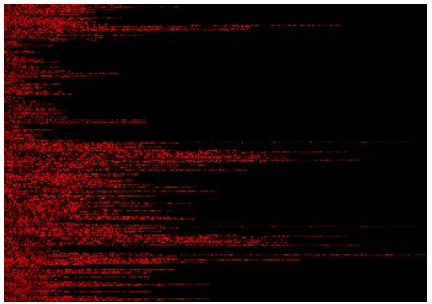
- Classify the data in the multiclass manner - divide the dataset into 7 classes (1 no cancer class and 6 cancer classes)

Datasets

Datasets

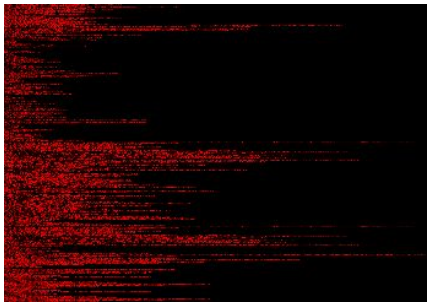
- Our datasets consist of 2 dimensional samples
- Each of them have 267 rows and 531 columns
- Rows represents signaling pathways
- Columns represents specific genes

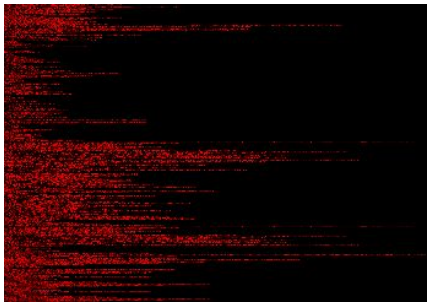




Datasets

- Signaling pathways belong to the 8 different biologically defined groups
- There are 24101 features (pixels) with non zero variance across train dataset
- Datasets are strongly imbalanced





I Dataset

This was our first dataset, split into three types cancer. We fit models on each cancer subset separately.

	Train set	***Test set***	***Imbalance ratio (Cancer vs NoCancer)***
OC	158	104	8.36
NSCLC	157	447	1.96
Sarcoma	118	56	1.8

II Dataset

In the next step we work onto bigger dataset with more types of cancer. In this part we fit model onto whole dataset.

Cancer vs NoCancer ratio = 1.38

	EC	***OC***	***NSCLC***	***GBM***	***Brain metastasis***	***Asymptomatic controls***	***Multiple sclerosis***
Train	39	28	142	215	25	260	65
Test	0	0	185	4	26	54	19
Total	39	28	327	219	51	314	84

III Dataset - multiclass variants

Finally we start working with the third dataset, which we use for the multiclass classification.

	Asymptomatic Controls	***Glioma and glioblastoma***	***NSCLC***	***Gastrointestinal***	***Gynecological***	***Neurological***	***Cardiovascular***
MultiGroup	405	128	567	318	171	126	201
MultiGroup2	405	128	567	489		126	201
MultiGroup3	405	128	567	489		327	

III Dataset - split

	Train	***Test***	***Cancer vs NoCancer ratio***
Split by hospitals	891	1025	2.25
Random split 70/30	1340	576	4.72

Used methods

Used methods

During our work we focused on the following types of machine learning algorithms:

- boosted trees
- convolutional neural networks
- MLP classifier build on the latent space of variational autoencoder

Boosted trees

Boosted trees

- We decided to use gradient boosted trees because it's an efficient algorithm in many applications
- We choose implementation from XGBoost library
- Another advantage is an ability to defined features importance

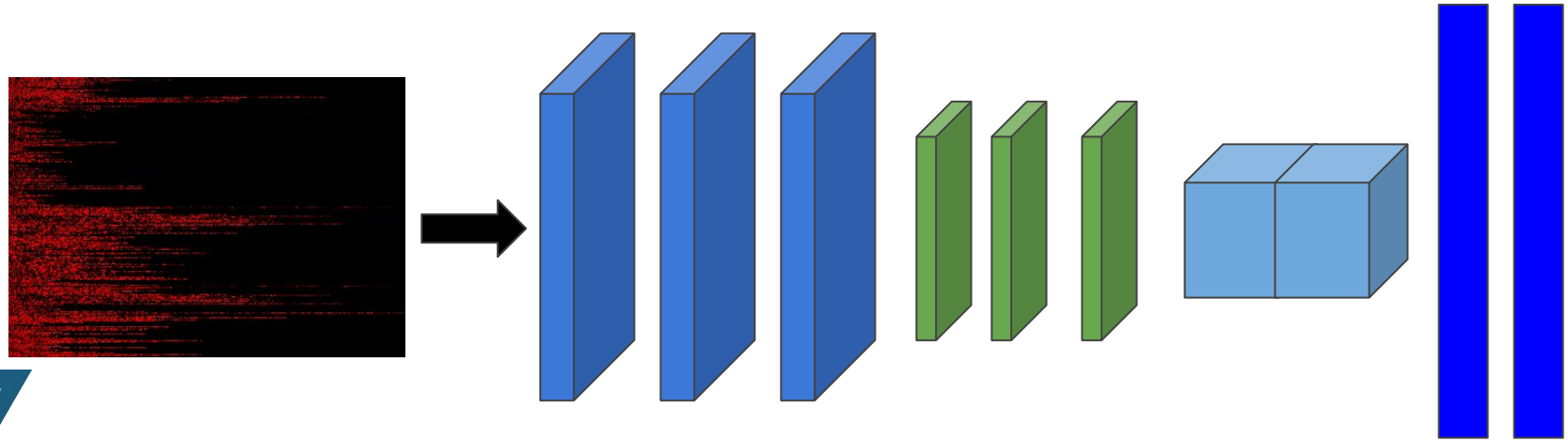
Convolutional neural networks

Convolutional neural networks

We decided to use CNNs because our data have originally shape of 2D array.

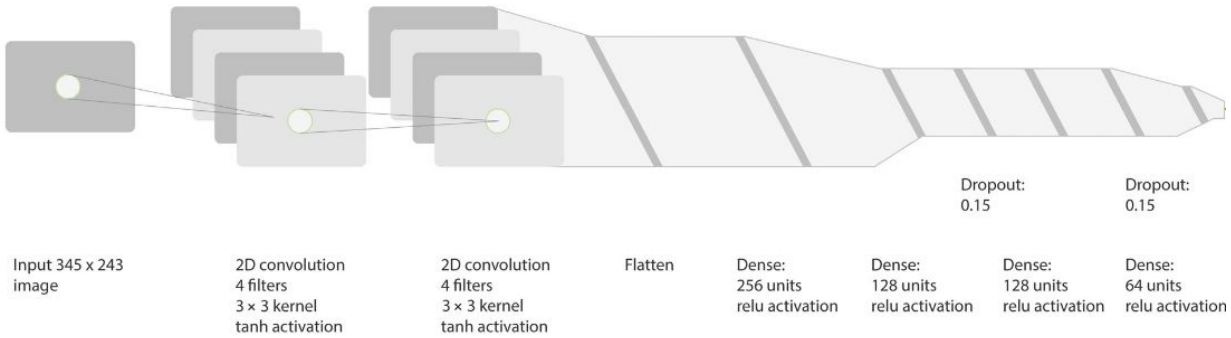
We tested two types of architecture:

- custom architecture called Implatelet
- standard architecture Resnet in 18 and 34 layers variants



Implatelet

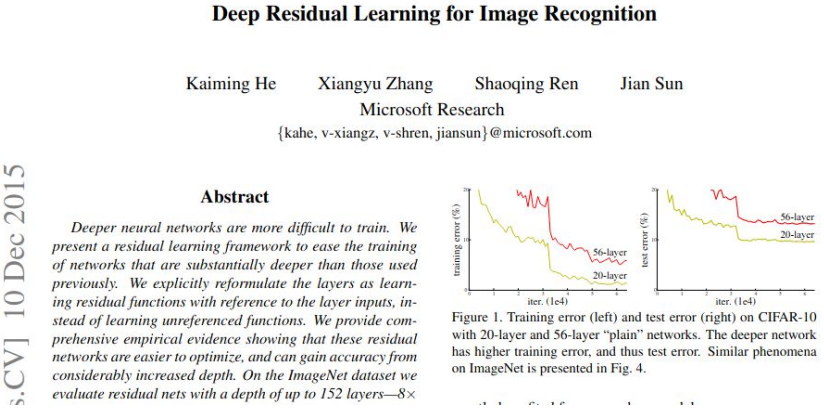
A custom architecture designed by Krzysztof Pastuszak for the specific task of cancer classification of liquid biopsy data. It contains > 90 mln of parameters.



K. Pastuszak et al., "implatelet classifier: image-converted rna biomarker profiles enable blood-based cancer diagnostics," Molecular Oncology, 2021

ResNet

- In this experiment we decided to use ResNet architecture in 18 layers variant
- We choose it due to it's balance between good results and quite small number of parameters in a model
- ~ 11 mln parameters vs 90 mln parameters of Implatelet



Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Abstract

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8×

Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

ResNet

Me made some modifications to the original version of model:

- add a Dropout layer before the output
- use mixUp data augmentation

We optimized parameters like:

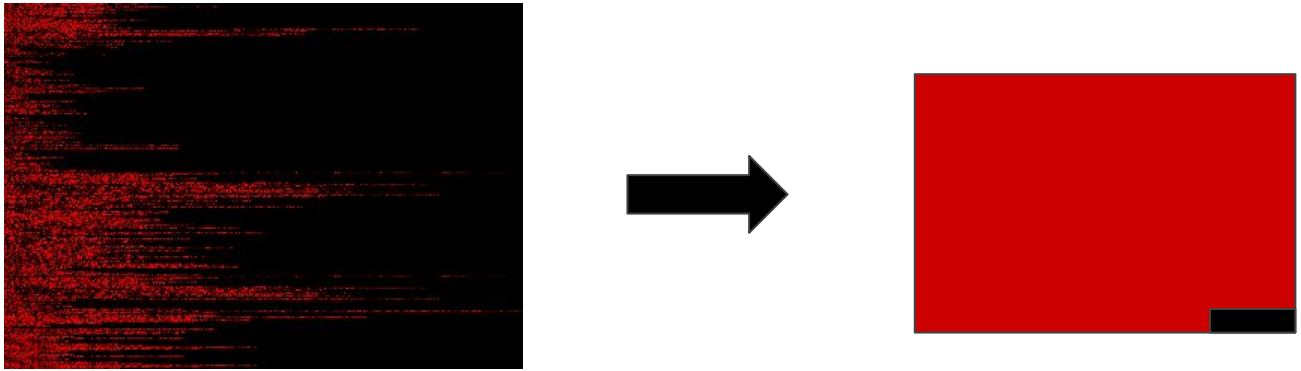
- learning rate
- dropout probability
- weight decay

```
num_ftrs = resnet.fc.in_features
# Here the size of each output sample is set to 2.
resnet.fc = nn.Sequential(
    nn.Dropout(dropout),
    nn.Linear(num_ftrs, 2)
)
```

Experiments

Standard vs reduced parametrization

- We reduce samples to pixels with non-zero variance and nonzero values.
- It results in reduction from 141 777 to 24 101 features.
- We form a new rectangle with shape 155 x 156



Experiments - Dataset I ovarian cancer

Backbone	***Validation bal. acc.***	***Test bal. acc.***
Standard parametrization		
ResNet18	0.9080	**0.8958**
ResNet34	0.8793	0.8317
Reduced parametrization		
ResNet18	0.8938	**0.8563**
ResNet34	0.9218	0.8255

mixUp augmentation

Data augmentation method proposed in: *Zhang, H. et al.: mixup: Beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018*

which can be expressed with given formula:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

where (x_i, y_i) and (x_j, y_j) are randomly selected training pairs of input vectors and the corresponding label, and $\lambda \in [0, 1]$ is the interpolating factor.

Experiments - Dataset I ovarian cancer tricks

Model	*Validation bal. acc.*	*Test bal. acc.*	*Test std. (3 train.)*
Resnet18			
ImageNet	0.9236	0.8952	0.0056
mixUp	0.9379	0.8798	0.0242
mixUp + ImageNet	0.9343	**0.9043**	0.0328
Resnet34			
ImageNet	0.9236	0.8652	0.0229
mixUp	0.9042	0.8221	0.0477
mixUp + ImageNet	0.9343	**0.8782**	0.0185

Experiments - full Dataset I

Cancer subset	*Validation bal. acc.*	*Test bal. acc.*
Boosted trees		
OC	1.0	0.8991
NSCLC	0.76	0.7343
Sarcoma	0.9818	0.6316
ResNet-18		
OC	0.9343	**0.9043**
NSCLC	0.9129	**0.8652**
Sarcoma	1.00	**0.9409**

Experiments – Dataset II – transfer to new hospital

Model	*Validation bal. acc.*	*Test bal. acc.*	*Validation AUC*	*Test AUC*
Boosting	0.909	**0.878**	0.967	0.953
ResNet-18	0.913	0.857	0.965	0.958
imPlatelet	0.898	0.854	0.970	**0.966**

II Dataset - feature importance

- We can see that just 100 parameters is enough for XGBoost trees to reach a 95% AUC on test set

- It's another step in feature reduction: 141 777 -> 24101 -> 100!

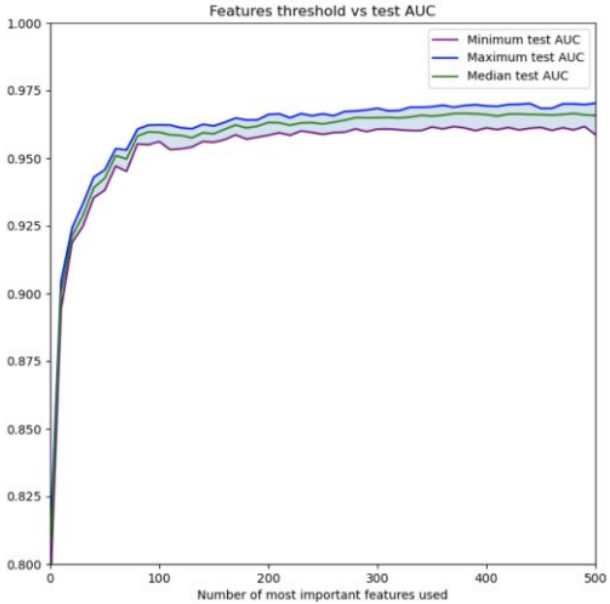


Figure from: Cygerts, et al. Platelet-based liquid biopsies through the lens of machine learning

Boosted trees experiments - Dataset III multiclass

Dataset	*Train AUC*	*Train Bal. acc.*	*Test AUC*	*Test Bal. acc.*
MutliGroup2				
RandomSplit	0.994	0.914	**0.784**	**0.428**
HospitalSplit	0.999	0.913	0.715	0.402
MutliGroup3				
RandomSplit	0.999	0.993	**0.811**	**0.530**
HospitalSplit	1.0	1.0	0.705	0.426

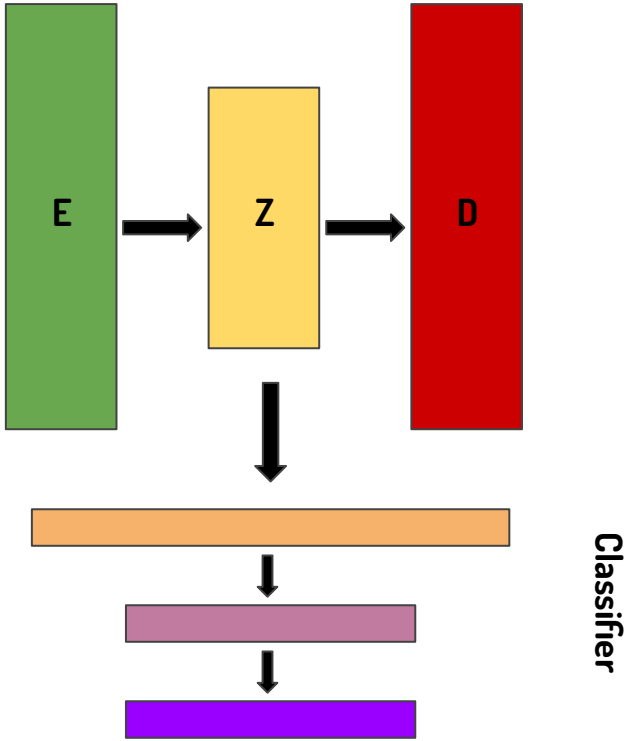
CNN experiments – Dataset III multiclass

Dataset	***Train AUC***	***Train Bal. acc.***	***Test AUC***	***Test Bal. acc.***
MutliGroup2				
RandomSplit	1.000	1.000	**0.839**	**0.481**
HospitalSplit	0.986	0.812	0.663	0.282
MutliGroup3				
RandomSplit	1.000	1.000	**0.864**	**0.588**
HospitalSplit	0.999	0.837	0.667	0.351

VAE + Classifier

VAE + Classifier

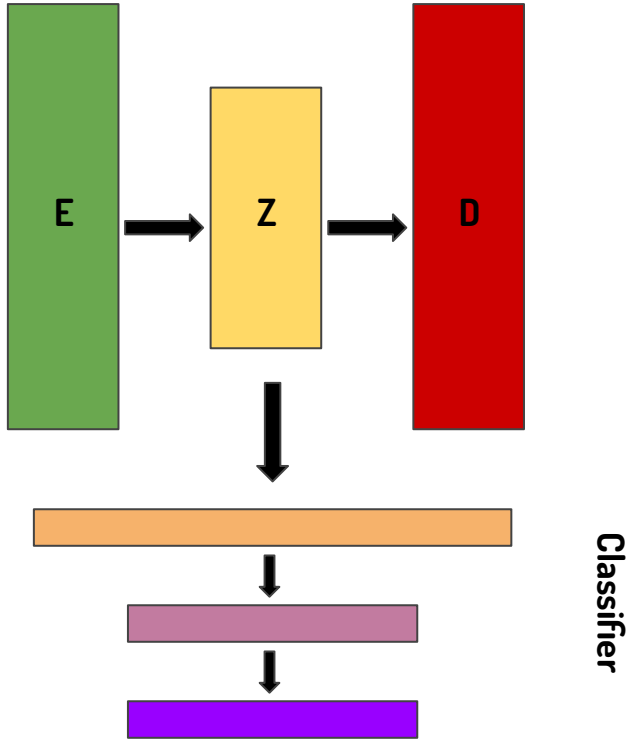
- Using Autoencoder as a features reduction method
- New data representation is from laten space *z*
- There is simple MLP classifier build on a latent space



Zhang, et al. Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

VAE + Classifier

- We can reduce samples dimensionality from 24 101 -> 128
- We can overfit our model on training data and reach ~ 100% AUC and bal. acc. for multiclass classification
- Now we struggle with regularization of a model for getting better test results - currently ~ 70% AUC and ~40% bal. acc.



Zhang, et al. Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Thank you!
That's all

Questions
&
Discussion