A case study for the identification of cancer stages in breast cancer

Garali[13], Renault[13] and Deleuze[123]

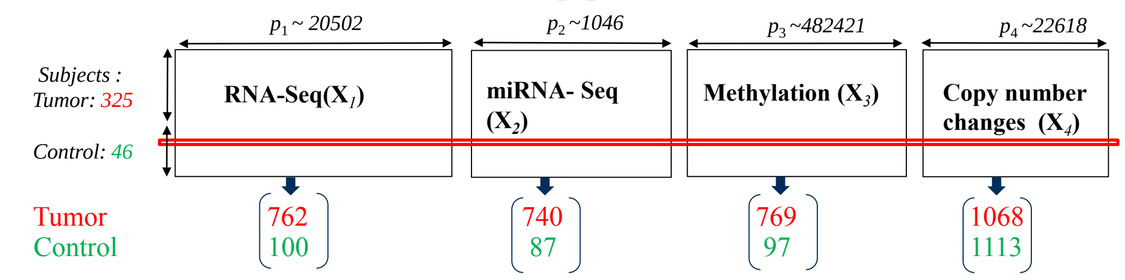
[1]Fondation Jean Dausset-CEPH (Centre d'Etude du Polymorphisme Humain), Paris, France; [2]Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France; [3]The Laboratory of Excellence in Medical Genomics, GENMED.

INTRODUCTION

The growing number of multi-omics data, characterizing a given disease, provides physicians and statisticians with complementary facets of the disease process. However, novel statistical methods of data analysis are needed to unify these views. In order to confirm the expected richness of multi-dimensional data, we first tested deep learning approach on one single data type, RNA-Seq data, to predict breast cancer stages. Secondly, deep learning results of RNA-Seq data are compared to traditional machine learning techniques. Finally, a comparative result with integrative analysis method is presented.

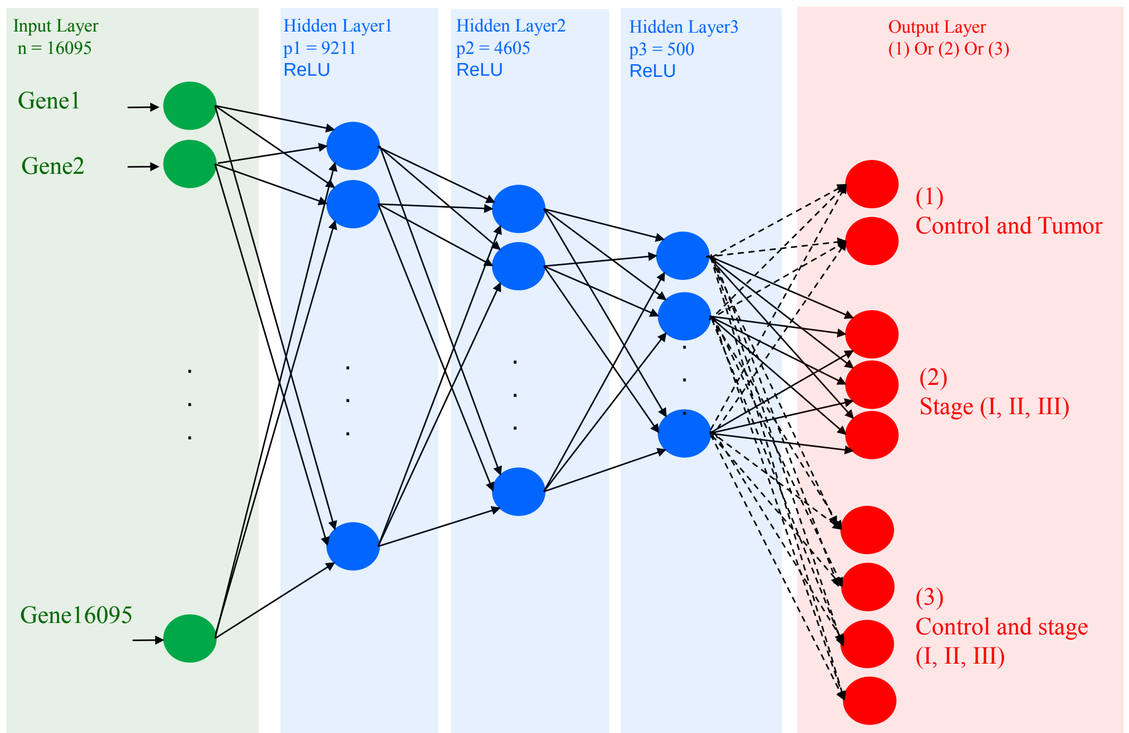
I. DATASET: MULTI-OMIC

Multi- Omics data from TCGA data: BRCA [1]



II. DEEP LEARNING MLP AND RNA-SEQ DATA

II. 1) MLP: (Multi Layer Perceptron)



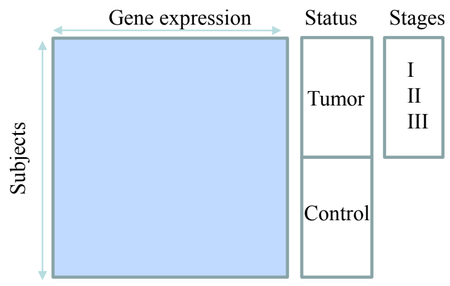
Problem	Type Last-layer activation	Loss function
(1): Binary classification	sigmoid	binary_crossentropy
(2) and (3): Multiclass, single-label classification	softmax	categorical_crossentropy

II. 2) RNA-Seq Data

RNA-Seq data:
- ≈ 16095 genes after filtering

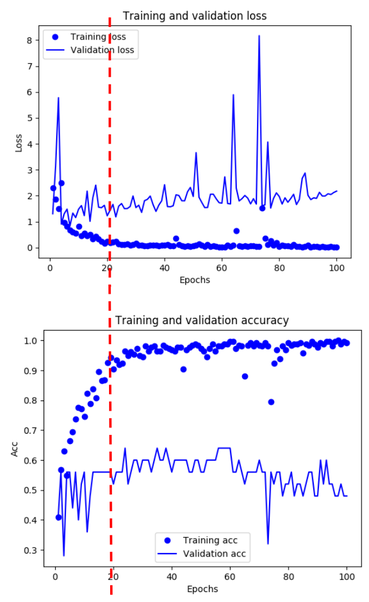
- 424 samples:
  - 100 control
  - 324 tumor

- Tumor sample:
  52 Stage I, 195 Stage II and 77 Stage III

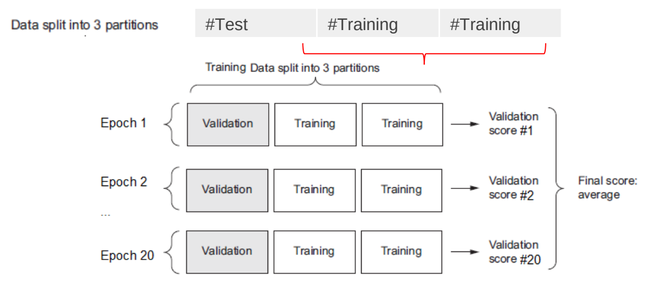


II. 3) Avoiding Overfitting

-Regularization technique: dropout (0,5)
-Early stopping: 20 Epoch



II. 4) Validation



II. 5) Results

	Network1	Network2	Network3	Network4	Network5	Network6	Network7	Network8
Hidden layer (HL) and Neurones	HL1:9211 HL2:4605 HL3:500	HL1:9211	HL1:9211 HL2:500	HL1:1000 HL2:200	HL1:1000	HL1:10000	HL1:200	HL1:9211 HL2:4605 HL3:500 HL4:200
Acc. MLP(1)	0,97	0,98	0,99	0,99	0,97	1	0,99	0,99
Acc. MLP(2)	0,61	0,61	0,57	0,53	0,55	0,57	0,50	0,61
Acc. MLP(3)	0,70	0,65	0,71	0,67	0,68	0,70	0,68	0,72

III. COMPARISON WITH DIFFERENT STUDIES

III. 1) Traditional machine learning techniques

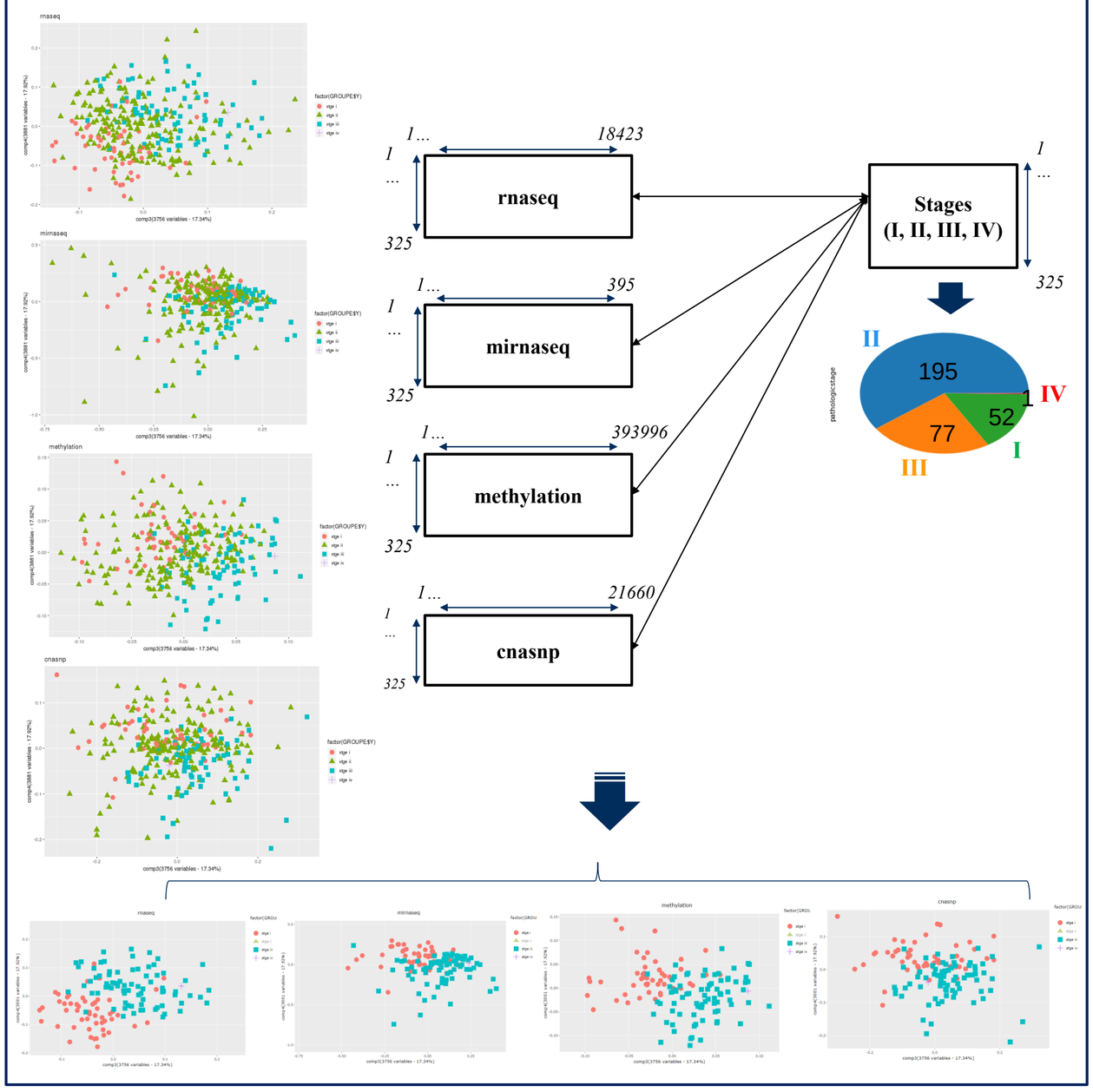
Method	Logistic Regression	Nearest Neighbor	SVM(linear)	SVM(rbf)	Naïve Bayes	Auto-keras	MLP(2)	Decision Tree Algorithm	Random Forest Classification
Acc. (Stages)	0,34	0,39	0,54	0,62	0,58	0,62 (15 models)	0,61	0,80	0,81

III. 2) Regularized Generalized Canonical Correlation Analysis (RGCCA) [2]

1-RGCCA can process a priori information defining which blocks are supposed to be linked to one another, thus reflecting hypotheses about the biology underlying the data blocks.
2-RGCCA integrates a variable selection procedure, called SGCCA, allowing the identification of the most relevant features.
3-The RGCCA/SGCCA-based integrative analysis method aims at summarizing the relevant information between and within the blocks.
4-The introduction of the design matrix $\boldsymbol{C}$, the shrinkage parameters τ_j and the scheme function g makes RGCCA (1) highly versatile.

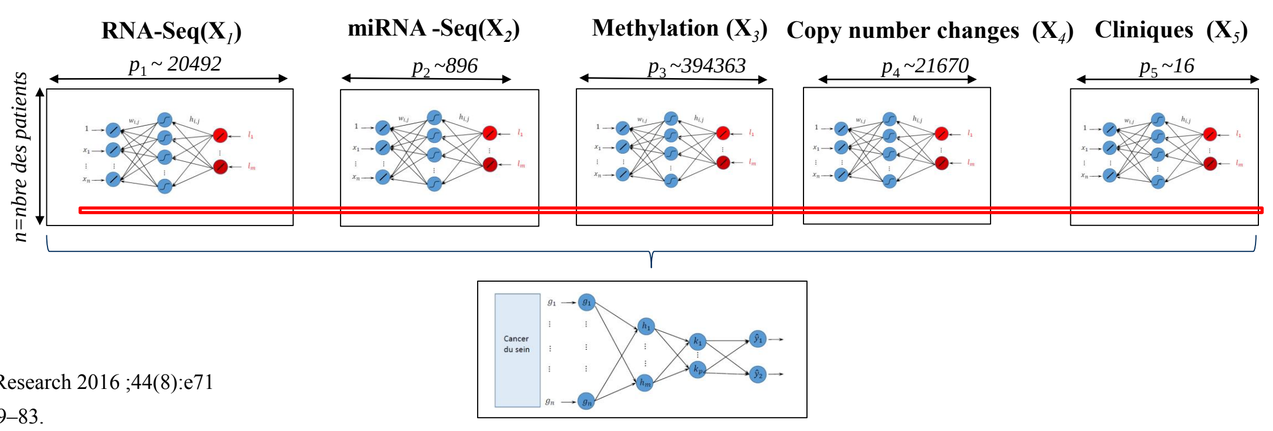
$$(1)\quad \max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk}\, g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j,\ \mathbf{X}_k\mathbf{w}_k\right)\right)$$

$$\text{s.t.}\ \left(1-\tau_j\right)\mathrm{var}\left(\mathbf{X}_j\mathbf{w}_j\right)+\tau_j\left\|\mathbf{w}_j\right\|_2^2 = 1\ ,\ j=1,\ \ldots,\ J$$



CONCLUSION

In this work, we compared traditional machine learning techniques to deep learning models for the identification of breast cancer stages. Then, we used integrative analysis method, RGCCA/SGCCA. Our results show that the benefit of using deep learning models remains unclear. On the one hand, deep learning models suffer from instability and overfitting. On the other hand, using the various genomic data, multimodal fusion should improve classification rates. Thus there is a need to develop a multimodal method for breast cancer stages prediction, to identify a selection of subset of genes as a signature.



[1] Colaprico A, Silva T.C, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research 2016 ;44(8):e71

[2] Tenenhaus A, Philippe C, Guillemot V, et al. Variable selection for generalized canonical correlation analysis. Biostatistics 2014;15(3):569–83.