# Research Project Proposal: Machine Learning algorithms applied on RNA-seq data to classify subtypes of Colorectal Cancer

CHIARA BARBERA, CHIARA.BARBERA@MAIL.POLIMI.IT

## 1. Introduction to the problem [max 1 page]

This document is intended to describe what kind of research we will perform and how it will be performed. The area of our work is the classification of subtypes of colorectal cancer (CRC). This subject involves both genomics, computation and computer science, since it regards the analysis of gene expression data and the execution of machine learning algorithms. This research topic is valuable because the knowledge of the subtype of CRC can help improving the diagnosis and the therapy of CRC. Even if we will start performing a replication of the work of Isella *et al.* [1] that introduced the CRIS classification (ColoRectal Intrinsic Subtype), **our main focus will be to develop a single-sample classifier starting from the one they proposed, possibly testing alternative classifiers**: the ultimate goal is **improving the current performances and reach higher accuracy and robustness** with respect to data production pipelines and normalization procedures. In the development of the new classifier, we will have to face some issues, including:

**Dataset size and unbalance**: we will cope with a relatively small number of samples; moreover, because of their biological origin, the samples may not be evenly distributed among the subtypes, naturally reflecting their diffusion in the population. Because of these two reasons, we will have to carefully select how to distribute and balance the samples between and within training and testing: a carefully balanced cross-validation approach, for example, would allow to efficiently exploit a small dataset, since the samples would be made turn around between training and validation set. An unbalanced dataset will also influence the metrics used to evaluate the classifier performances (See Section 3 for further information on metrics).

**Curse of dimensionality**: first introduced by Bellman in [2], it refers to the exponential growth of needed data to estimate a function with multiple features. In our topic, we will deal both with a huge variety of genes and a limited number of sample data. Because of this, we will have to carefully select the most relevant genes, for example through well-designed feature selection algorithms. Even if this type of procedure has been already run in [1], we will evaluate it since we will work on a different type of data (NGS RNA-seq data).

**Preprocessing and normalization bias**: since we are working with biological data (which are intrinsically noisy), the normalization and the pipelines used to preprocess the data may introduce some bias that influences the classification.

Differently from [1], we will consider **only RNA-seq data, possibly investigating different types of normalization.** The usage of a unique and specific type of data and of a uniform normalization method will make the results of the evaluation easily comparable. We chose to consider RNA-seq data because of the advantages offered by the NGS (Next Generation Sequencing) technique, like the availability of more potentially relevant features (with respect to microarray data) and the higher quantification accuracy. This may lead to discover new features that are relevant for CRC cancer subtyping, possibly providing us with the means of even defining a new classification with better performance.

## 2. Main related works

The main related work for us is the research done by Isella *et al.* [1]: they improved the previous standard classification system, the Consensus Molecular Subtype (CMS) [3], which was influenced by the stromal content [1]. Isella *et al.* identified 5 ColoRectal Intrinsic Subtypes (CRIS) and tried to implement also a single-sample classifier, whose performances though were not optimal. Other relevant works include the research of Franks *et al.* [5], who developed a new normalization procedure that removes the influence from the platform with which the data have been collected. Also, Kim *et al.* [6] tested several machine learning algorithms to distinguish between normal vs. cancerous cells and between 21 types of cancer. Finally, in 2019, Murcia *et al.* [7] validated another classification system based on CIMP [2], MSI, [3] mutations of BRAF and KRAS genes and studied the influence on response to chemotherapy; however, they used the classification developed by Phipps *et al.* [8] and Sincrope *et al.* [9], which were published even before the CMS. Because of this, we will start from the work of Isella *et al.*, which can, in our knowledge, be considered a state of the art classification system for CRC subtypes.

## 3. Research plan

Our main goal is the implementation of different machine learning classifiers for CRIS subtypes [1] and possibly the development of a new classification based on RNA-seq data. Our research will be predominantly hybrid: its experimental side lies in the investigation of a novel type of classification and in the attempt for the first time, to our knowledge, to apply alternative classifiers to the CRIS classification. However, our research has also an application side because its results may be translated, in future, into a clinical application for the prediction of CRC prognosis and drug sensitivity. Our work will follow **six phases**, whose details are represented in the Gantt diagram in Figure 1.

1. **Set up and state of the art**: after having understood the problem and performed a research of the state of the art, we will collect and preprocess all the gene expression data. In particular, we will use RNA-seq data of CRC.

2. **Replication and assessment of CRIS classifier**: we will study in deep the NTP and the k-TSP classifiers proposed in [1] and execute them on the RNA-seq data only. We will consider different normalization procedures and different dataset composition to test the robustness of the classifiers. We will compare the results with the performances of the previous NTP, k-TSP and with the previous state of the art classification (the Consensus Molecular Subtypes, [3]).

3. **Alternative classifiers**: we will evaluate different classification algorithms to improve the existing single-sample k-TSP classifier. We will select the relevant features and samples according to biological and computational criteria (e.g. performing feature selection, checking if the samples are annotated for the needed genes) and execute the classifiers, choosing the one with the best performances.

4. **Comparative analysis**: we will compare the performances of all the classifiers that we executed in the previous phases.

5. **Clinical and biological validation**:We will perform two types of analysis, the Gene Set Enrichment Analysis (GSEA) [10] and the Sample Set Enrichment Analysis (SSEA)[1], to identify, among the genes that we selected as relevant, which of them are associated with drug sensitivity and prognosis prediction through the interpretation of their expression.

6. **Paper writing**: we will finalize the writing of a paper with the data collected through periodic reports (written at the end of each phase).

[1] **Stroma**: tissue composed of cells that serve as a matrix in which the other (tumoral) cells are embedded [4].
[2] **CIMP** = CpG Island Methylator Phenotype.
[3] **MSI** = MicroSatellite Instability.

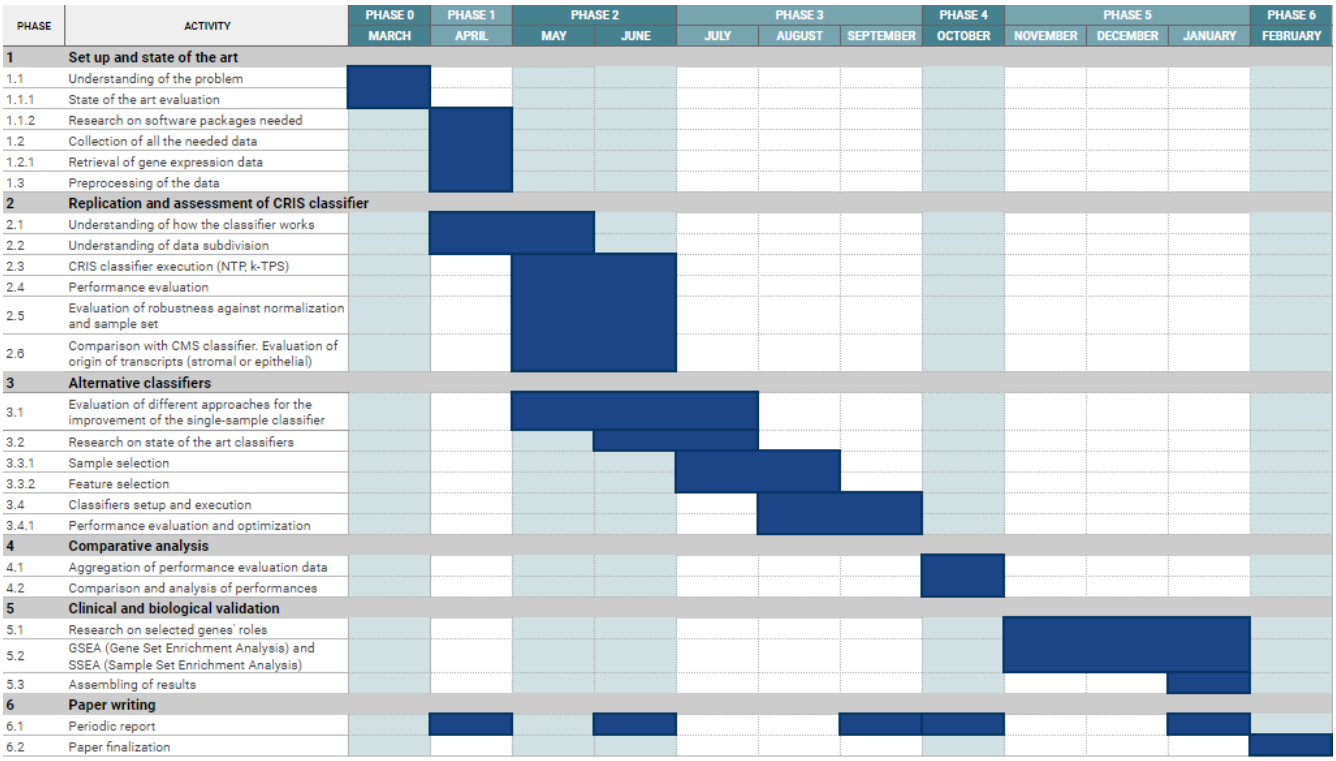


*Figure 1:* ***Gantt diagram*** *for our project, which should be concluded by February 2021.*

We conclude this section with a list of candidate metrics for the evaluation of the classifiers. We will consider also other global metrics for multiclass settings, like the **microaverage** and the **macroaverage** [11] versions of the following metrics:

1. **Confusion matrix**: displays, for each class, the number of correctly classified samples and the number of misclassified samples, comparing the results of a classifier with respect to a target reference assumed as true. For each class, true positives and true negatives represent the samples correctly classified as belonging (or not belonging) to the class, respectively. Since the CRIS classification has defined five classes, the confusion matrix will be a 5x5 square matrix with elements $x_{ij}$ (number of samples of class $i$ that have been classified by the algorithm as belonging to class $j$).
2. **Precision**: represents the positive predictive value [12].
3. **Recall (sensitivity)**: represents the true positive rate [12].
4. **F1-score**: the armonic mean of precision $P$ and recall $R$.
5. **Accuracy**: for 5 classes, it is defined as the number of correctly classified samples within all the classes with respect to the total number of samples. We will consider also the **balanced accuracy**, which is the average of the recalls of the single classes.
6. **Specificity**: represents the true negatives rate [12].
7. **Matthew Correlation Coefficient** (MCC)[13]: it is more representative with respect to the F1-score in cases where the number of samples for each class is unbalanced. MCC ranges from -1 (disagreement between prediction and observation) to 1 (perfect prediction).

## References


[1] C. Isella, F. Brundu, S. E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, C. Boccaccio, E. Ficarra, L. Marchionni, L. Trusolino, E. Medico, and A. Bertotti, "Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer," *Nature Communications*, vol. 8, p. 15107, 2017.

[2] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[3] J. Guinney, R. Dienstmann, X. Wang, A. Reyniès, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, B. Bot, J. Morris, I. Simon, S. Gerster, E. Fessler, F. De Sousa E Melo, E. Missiaglia, H. Ramay, D. Barras, and S. Tejpar, "The consensus molecular subtypes of colorectal cancer," *Nature Medicine*, vol. 21, pp. 1350–1356, 2015.

[4] "Tissue | definitions, types & facts." `https://www.britannica.com/science/tissue#ref163759`.

[5] J. Franks, G. Cai, and M. Whitfield, "Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data," *Bioinformatics (Oxford, England)*, vol. 34, p. 1868–1874, 2018.

[6] B.-H. Kim, K. Yu, and P. Lee, "Cancer classification of single cell gene expression data by neural network," *Bioinformatics (Oxford, England)*, vol. 36, p. 1–7, 2019.

[7] O. Murcia, M. Juárez, M. Rodríguez-Soler, E. Hernández-Illán, M. Giner-Calabuig, M. Alustiza, C. Egoavil, A. Castillejo, C. Alenda, V. Barberá, C. Mangas-Sanjuan, A. Yuste, L. Bujanda, J. Clofent, M. Andreu, A. Castells, X. Llor, P. Zapater, and R. Jover, "Colorectal cancer molecular classification using braf, kras, microsatellite instability and cimp status: Prognostic implications and response to chemotherapy," *Plos One*, vol. 13, p. e0203051, 2018.

[8] A. Phipps, P. Limburg, J. Baron, A. Burnett-Hartman, D. Weisenberger, P. Laird, F. Sinicrope, C. Rosty, D. Buchanan, J. Potter, and P. Newcomb, "Association between molecular subtypes of colorectal cancer and patient survival," *Gastroenterology*, vol. 148, pp. 77–87, 2015.

[9] F. Sinicrope, Q. Shi, T. Smyrk, S. Thibodeau, R. Dienstmann, J. Guinney, B. Bot, S. Tejpar, M. Delorenzi, R. Goldberg, M. Mahoney, D. Sargent, and S. Alberts, "Molecular markers identify subtypes of stage iii colon cancer associated with patient outcomes," *Gastroenterology*, vol. 148, pp. 88–99, 2015.

[10] A. Subramanian, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. proc natl acad sci u s a," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 15545–15550, 10 2005.

[11] "Performance measures for multi-class problems." `https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/`.

[12] K. M. Ting, *Confusion Matrix*, pp. 209–209. Boston, MA: Springer US, 2010.

[13] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.