Statistical Learning Methods for Colorectal Cancer Signature Reconstruction and Classification in Patients with Chronic Inflammatory Bowel Disease

Mariem Abaach and Ian Morilla

December 7, 2021

1 Supplementary Methods

1.1 Data Normalisation

We use the standard Robust Multi-array Average (RMA) procedure, which was primarily designed for analysing gene expression data from Affymetrix arrays to normalise our samples. Results yielded before and after the normalisation are displayed in Fig. 1a to the whole sample and 1b to only the MIMAT miRNAs respectively. The corrected Gaussian distributions are well-centred around the median and present a proper kurtosis removing the unwanted variation of the original data. This normalisation enables a proper later $log2$ transformation to better detect the small differences among measures. The most right bottom panels in Fig. 1a and 1b still exhibit a heterogenous expression affecting the sensitivity of the Euclidean-based measurements as shown in the next section.

1.2 Unsupervised Hierarchical Clustering of Patients

In Fig. 2a we can observe the tree structure of the miRNA signature for all patients (CD and UC). Contrary to the expected tree CD and UC patients are disposed in two different branches; indeed, the tree was composed of three branches: first for CD cases, the second for UC cases, and the third for the control patients. As demonstrated below, pairwise comparisons based on $t - test$ were computed in order to cluster patients in an unsupervised hierarchical classification.

1.3 Patients Stratification

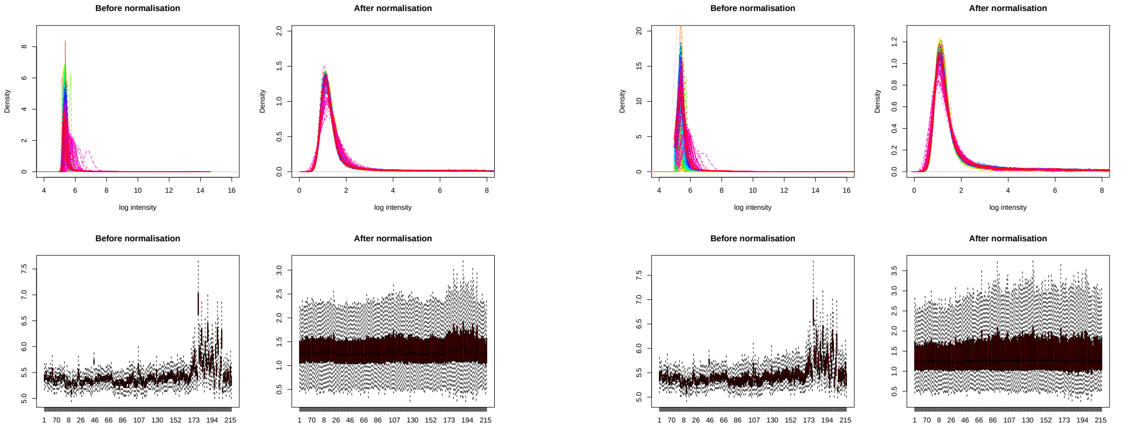
Fig. 2b explains how our cohort of patients non-linearly clusters before using any analysis that classifies them per strata. Details on the heatmaps with strategies 2 and 3 using the Euclidean norm are shown in Fig. 3. The results plotted in those panels demonstrate similar qualitative performances in the two cases.

1.4 Random Forests Setup

Herein we provide the description of the error propagation associated to random forests models developed for the entire sample, Crohn and UC patients respectively. We reach a steady error propagation as number of trees approaches 5000 Fig. 4. Regarding the performances of models CD learns much more accurately compared to ALL patients model being UC an intermediate case, but still displaying fair learning rates. On the other hand Fig. 5 shows error estimates, but using the strategies 2 and 3 respectively.

1.4.1 Variable Importance Analysis

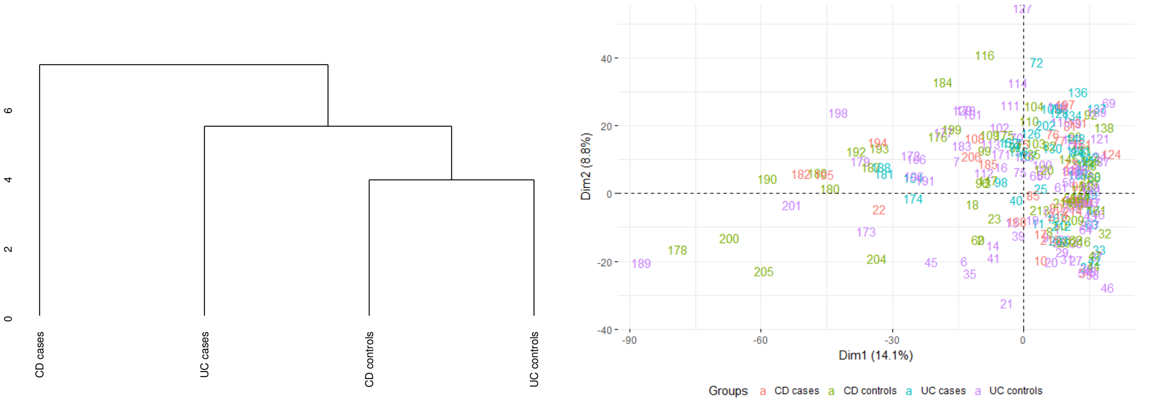
The plots in Fig. 6 explain VIMP of each group of patients for the RF models. Briefly, if a predictor is important in the current model, then assigning other values for that predictor randomly but "realistically" (i.e.: permuting this predictor's values over your dataset), should have a negative influence on prediction, i.e.: using the same model to predict from data that is the same except for the one variable, should give worse predictions.



(a) RMA normalisation of 216 miRNA candidates

(b) RMA normalisation of those MIMAT among the 216 miRNA candidates

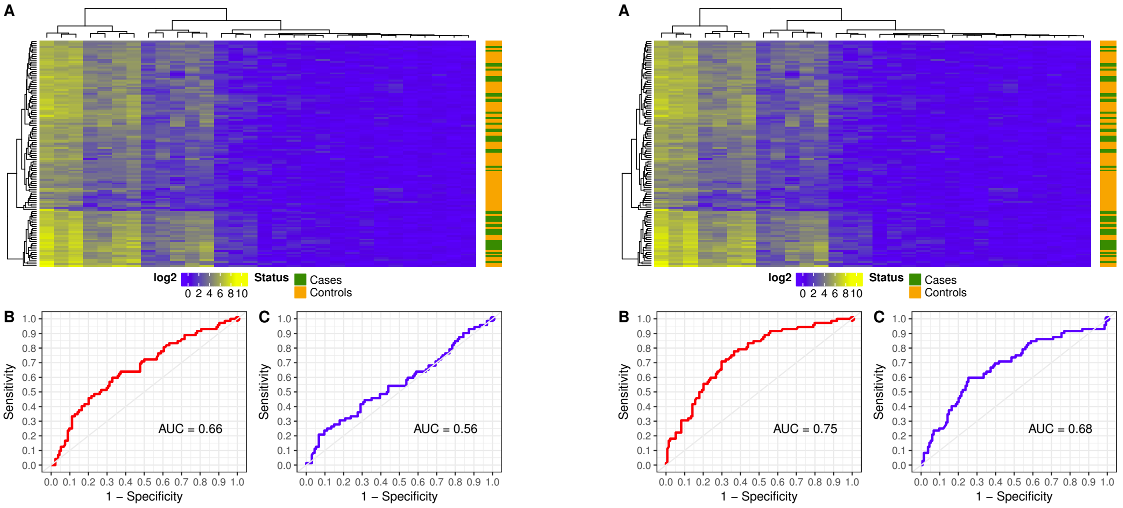
Figure 1: RMA normalisation of microarray data.



(a) Tree structure of the MIMAT miRNAs

(b) Strata of patients using PCA prior to use any analysis

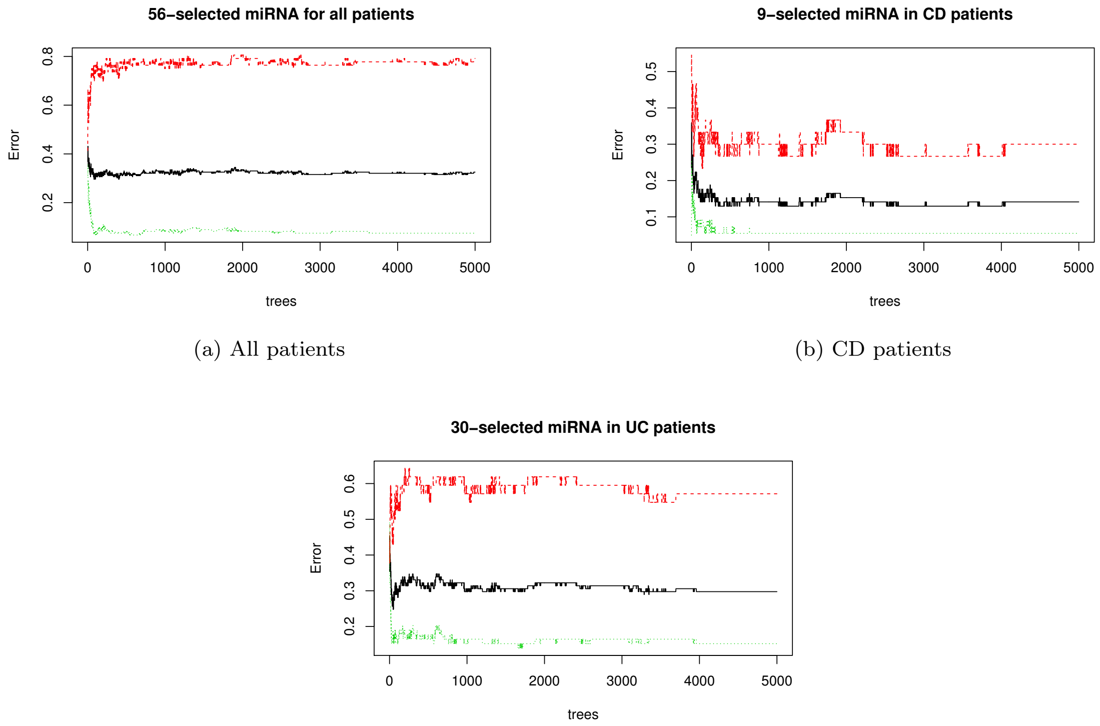
Figure 2: Tree structure of selected transcript miRNA and patients early stratification.



(a) Strategy 2

(b) Strategy 3

Figure 3: Unsupervised hierarchical clustering of all patients using strategies 2 and 3.

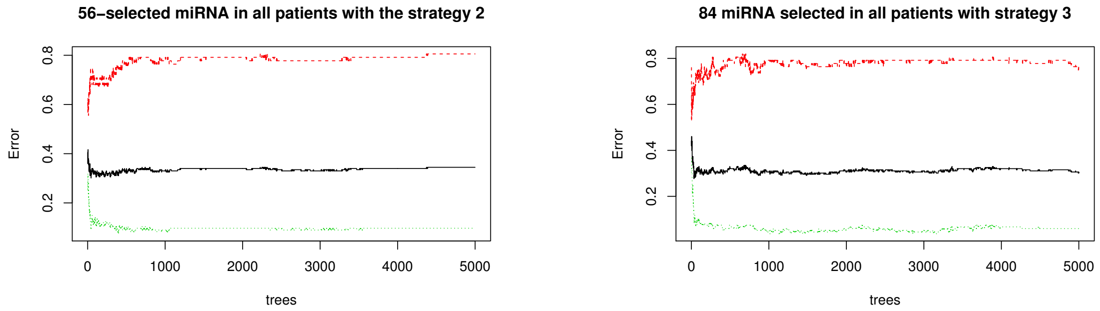


(a) All patients

(b) CD patients

(c) UC patients

Figure 4: Error propagation of our models: Random Forests with 5, 000 trees. Highlighted in red one may observe cases, in green controls and in black the overall error.



(a) Strategy 2

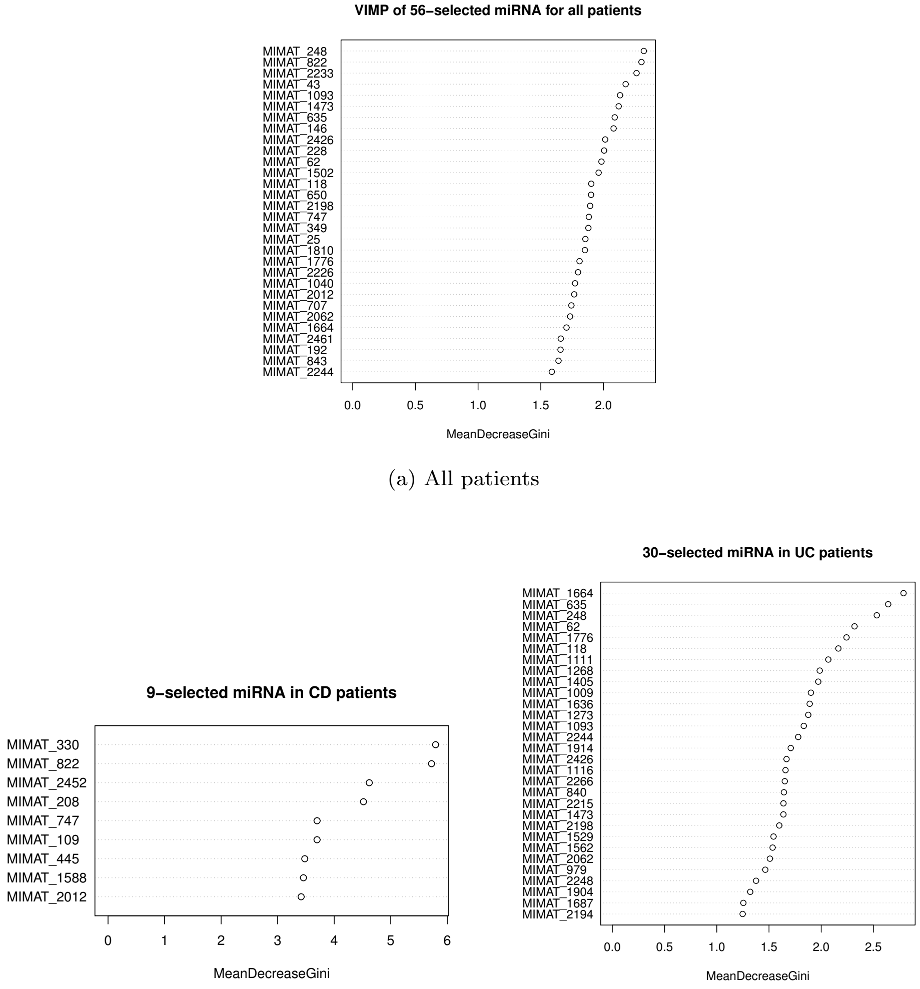
(b) Strategy 3

Figure 5: Error propagation of our models using strategies 2 and 3: Random Forests with 5, 000 trees. Highlighted in red one may observe cases, in green controls and in black the overall error.

So, we take a predictive measure (MSE) with the original dataset and then with the "permuted" dataset, and then we compare them somehow. One way, particularly since we expect the original MSE to always be smaller, the difference can be taken. Finally, for making the values comparable over variables, these are scaled.

1.5 sparse Partial Least Square Discriminant Analysis

In this section the effect of our methodology combined with sPLSD analysis is particularly evaluated. We introduce the results yielded by RF and SVM learning models in terms of their Roc curves Fig. 7a-7b as well as the study of their variable importance as plotted in Fig. 9.

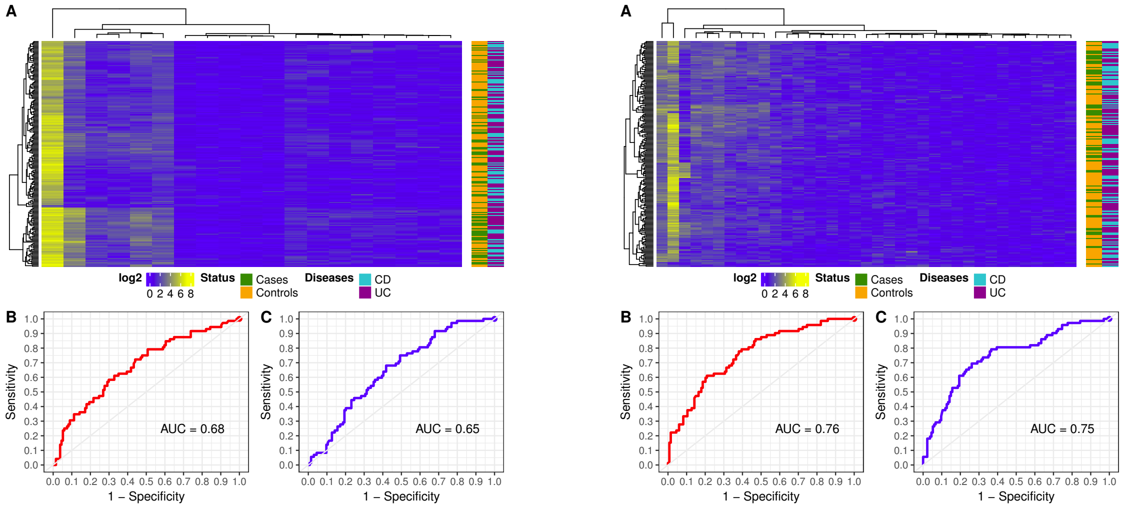


(a) All patients

(b) CD patients

(c) UC patients

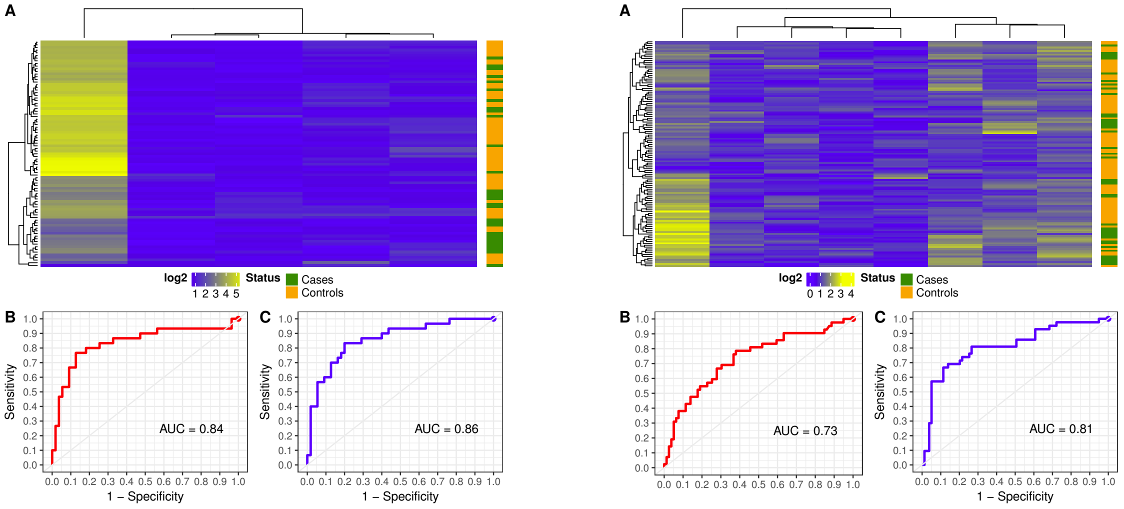
Figure 6: Variable Importance study of each Random Forests model. Due to the confidential nature of our miRNA signature, the MIMAT labels are mock transcripts and therefore they are not in accordance with the official miRNA symbols



(a) Strategy 2

(b) Strategy 3

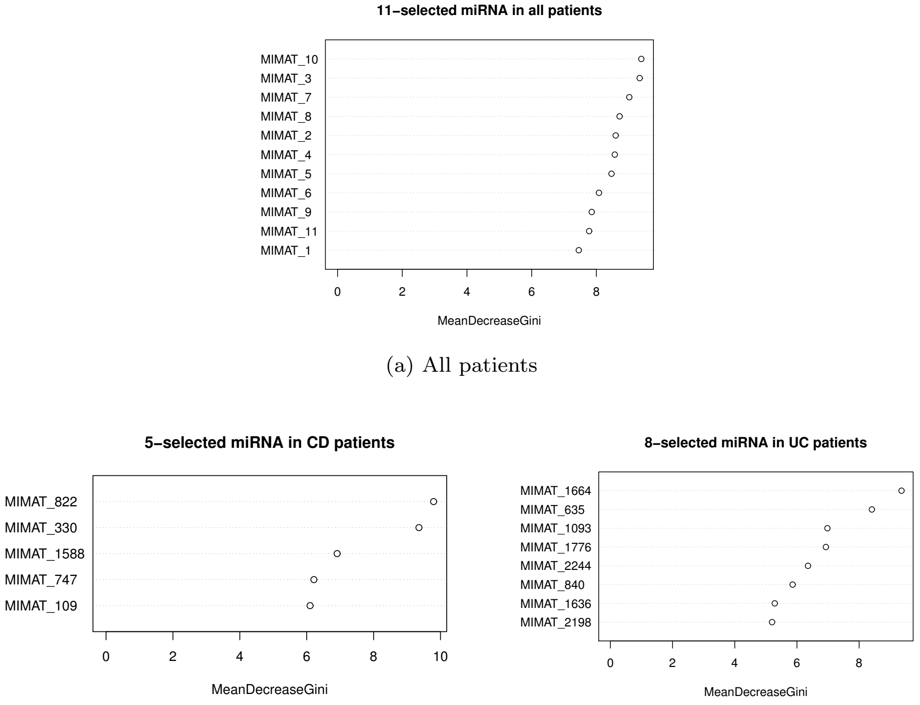
Figure 7: Unsupervised hierarchical clustering of all patients using strategies 2 and 3 combined with sPLSDA.



(a) Crohn patients

(b) Ulcerative Colitis patients

Figure 8: Unsupervised hierarchical clustering of Crohn and UC patients of the miRNA signature selected by sPLSDA.



(a) All patients

(b) CD patients

(c) UC patients

Figure 9: Variable Importance study of each Random Forests model. Due to the confidential nature of our miRNA signature, the MIMAT labels are mock transcripts and therefore they are not in accordance with the official miRNA symbols.