Towards an accurate cancer diagnosis modelization: Comparison of Random Forest strategies

Debit A[1,2], Poulet C[1], Josse C[1,4], Jerusalem G[4], Van Steen K[2], Bours V[1,3]

[1]University of Liege, GIGA-Research, Laboratory of Human Genetics, Liege, Belgium
[2] University of Liege, GIGA-Research, Medical Genomics, BIO3, Liege, Belgium
[3]University Hospital (CHU), Center of Human Genetics, Liege, Belgium
[4]University Hospital (CHU), Department of Medical Oncology, Liege, Belgium

LIÈGE université GIGA institute

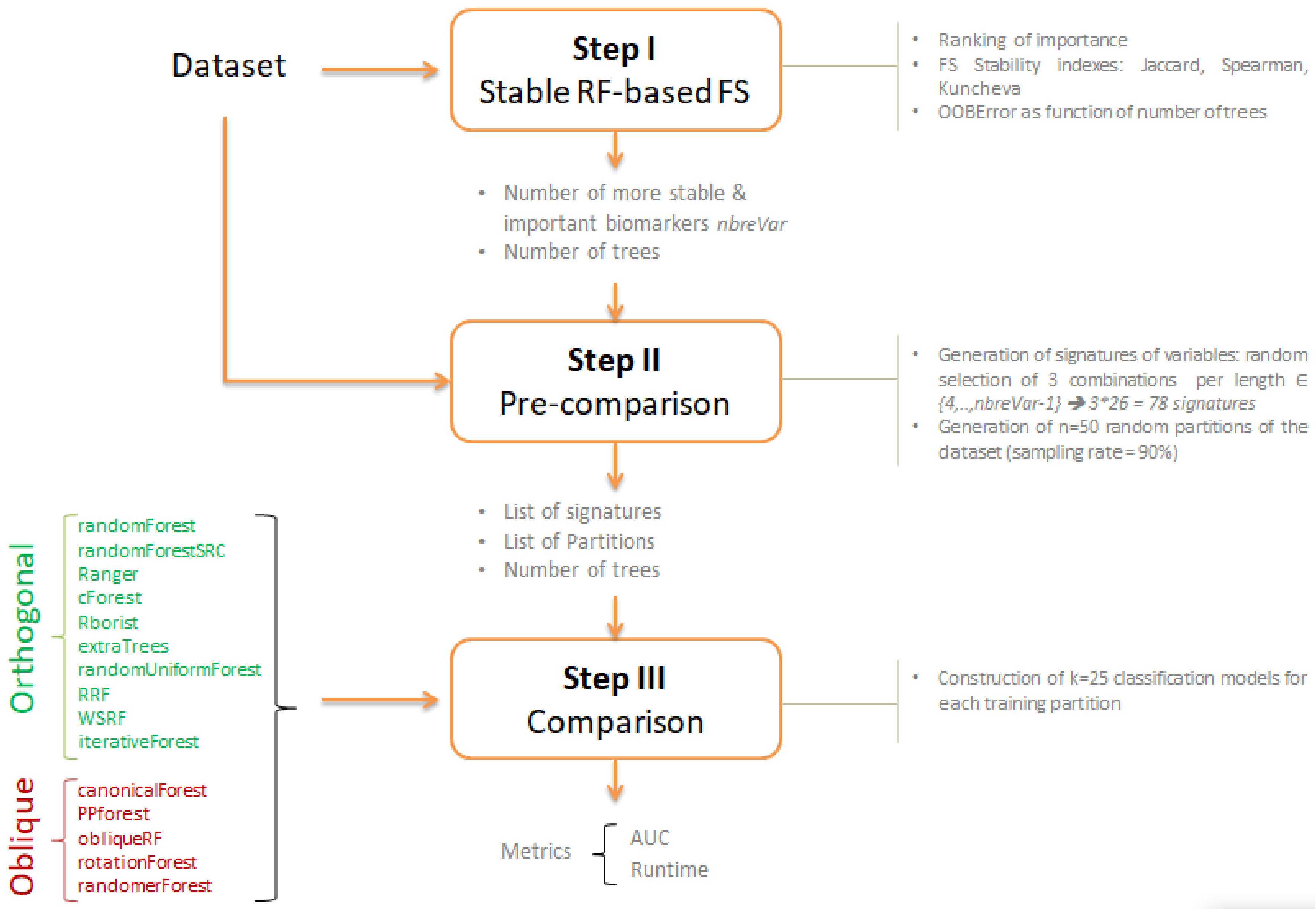
Background

Machine learning approaches are heavily used to produce models that will one day support clinical decisions. To be reliably used as a medical decision, such diagnosis and prognosis tools have to harbor a high-level of precision. Random Forests (RF) have been already used in cancer diagnosis, prognosis, and screening. Numerous Random Forests methods have been derived from the original random forest algorithm from Breiman et al. in 2001. Nevertheless, the precision of their generated models remains unknown when facing biological data. The precision of such models may be therefore too variable to produce models with the same accuracy of classification, making them useless in daily clinics.

Objectives

Empirical comparison of Random Forest based strategies, looking for their precision in model accuracy and overall computational time.

Differences/Similarities of the methods in the classification performance of the models built on different gene expression signatures



Dataset

Step I
Stable RF-based FS

- Ranking of importance
- FS Stability indexes: Jaccard, Spearman, Kuncheva
- OOBError as function of number of trees

- Number of more stable & important biomarkers *nbreVar*
- Number of trees

Step II
Pre-comparison

- Generation of signatures of variables: random selection of 3 combinations per length ∈ *{4,..,nbreVar-1} ➔ 3*26 = 78 signatures*
- Generation of n=50 random partitions of the dataset (sampling rate = 90%)

- List of signatures
- List of Partitions
- Number of trees

Orthogonal: randomForest, randomForestSRC, Ranger, cForest, Rborist, extraTrees, randomUniformForest, RRF, WSRF, iterativeForest

Oblique: canonicalForest, PPforest, obliqueRF, rotationForest, randomerForest

Step III
Comparison

- Construction of k=25 classification models for each training partition

Metrics: AUC, Runtime

Fig 1. Comparison pipeline of RF algorithms. step I: a stable feature selection (FS) is performed multiple times. The stability indexes are calculated over the resulted ranking sequences. The number of biomarkers maximizing the indexes is selected to generate the combinations of biomarkers. The number of trees needed to get a stabilized prediction error (OOB Error) is also calculated for each method. Step II: a random selection of 3 combinations per length is performed, and these partitions are used to build the models. Step III: A comparison is performed, involving the construction of k=25 models for each signature, each partition, and each method. The calculation of their performance (AUC), and computational times (modeling and prediction) are then assessed.

Materials and Methods

Main classification question

The difference between paired Tumor / Normal samples will be used as a strong classification parameter, allowing for strong modeling only.

Datasets

RNA-seq TCGA datasets:

- BRCA -Breast Invasive Carcinoma- (182 samples x 9560 genes)
- LUSC -Lung Squamous Cell Carcinoma- (96 samples x 9262 genes)

Random Forest based methods included in the comparison

Two families of methods: oblique and orthogonal, see figure 1 (bottom-left).

Experiments

- The experiments were carried out over 15 algorithms to classify paired normal-tumor patients from each dataset.
- Each of the selected gene expression signature is assessed for performance on each training partition and on each classification model for each algorithm ➔ In total: 50 x 25 = 1250 models for each signature for each algorithm
- All of the experiments were performed multiple times using the **same random training partitions** and the **same signatures**, everyone running on the **same computational nodes** for all the algorithms
- The R implementation of each algorithm was used
- The complete comparison protocol used in this study was displayed in figure 1.

Results

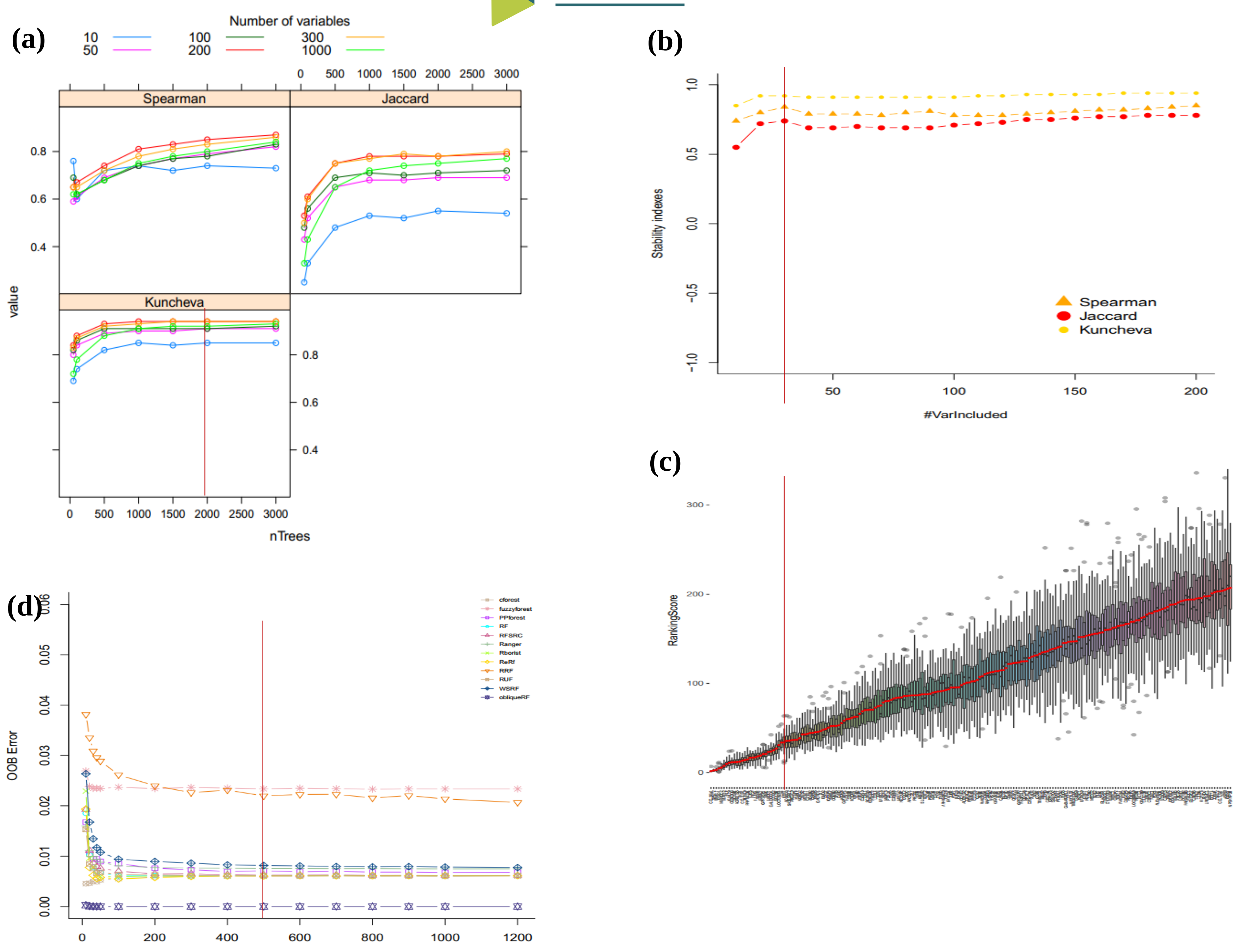


Fig 2. Results for the stability of the FS (BRCA-TCGA dataset): **a.** stability indexes for the feature selection are calculated for a different number of variables included in each ranking sequence and for an increasing amount of trees. The most stable feature selection is obtained using 2000 trees and 200 variables (vertical red line) . **b.** stability indexes using 2000 trees for the first 200 variables identifies the most important and most precise amount of important variables (30 for BRCA) indicated with red line. **c.** ranking distribution of the first 200 variables. **d.** prediction error (OOB error) calculated for the methods based on the selected variables coming from steps a, b, and c. The error is stable after 500 trees (red line) for all the methods.

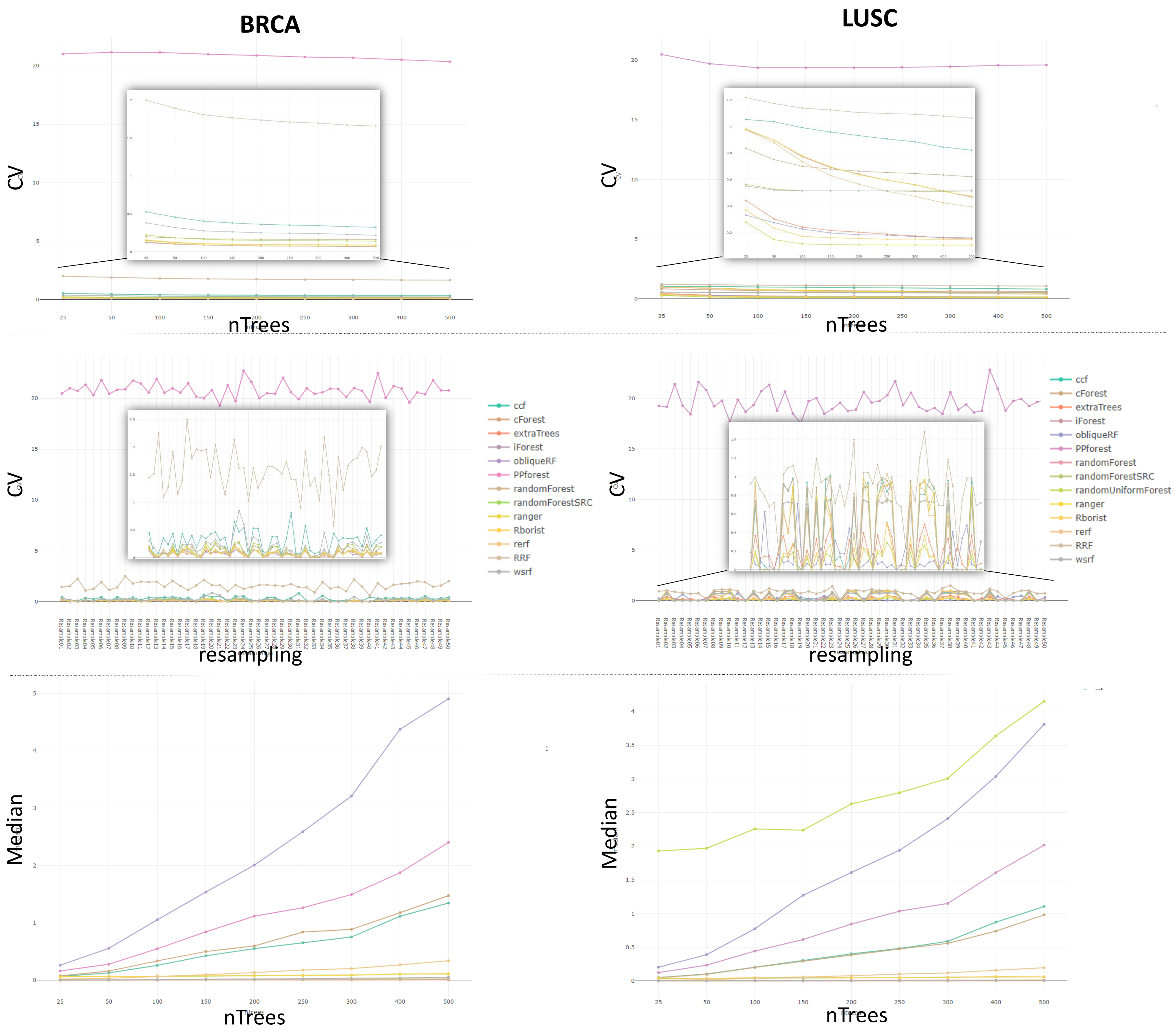


Fig 3 RF Comparison results. left: results for BRCA dataset, right: results for LUSC dataset. The precision of the models is approximated from the coefficient of variation CV (in percentage) of the AUC values. **a.** measure of CV as function of number of trees included in the models **b.** variation of CV over resampling. **c.** prediction runtime (median) in seconds

Discussion & Conclusion

Except for PPforest (CV > 19%), all RF methods are almost stable and accurate (CV ϵ [0,2]%), and dataset dependent

Some methods are more robust than others: PPforest (CV > 17), other methods (CV ϵ [0,2.5]%) ➔ resistance to sampling perturbation ➔ important in clinics

To select the best RF method for a given dataset, we take the trade-off between stability, robustness, and runtime.

Take home message:

Each algorithm should be tested over new datasets for their precision.

LIÈGE université GIGA institute · RÉGION WALLONNE · fnrs LA LIBERTÉ DE CHERCHER · cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Email: adebit@uliege.be