User-friendly and machine learning-empowered platform for classification of NSCLC based on RNA-seq profiling

Francois Collin[1], M. Niemira[2], A. Krętowski[2,3], J. Nikliński[4], M. Kwaśniewski[1]
Medical University of Bialystok, Poland: [1]Centre for Bioinformatics and Data Analysis, [2]Center of Clinical Research, [3]Department of Endocrinology, Diabetology and Internal Medicine, [4]Department of Clinical Molecular Biology
francois.collin@umb.edu.pl

Introduction

RNA-seq is a standard method for transcriptome profiling of cancer. Studies attempted to identify limited number of marker genes for diagnostic/prognostic of Non-Small Cell Lung Cancer NSCLC.

We propose to:
- use the complete transcriptome profile for **high prediction accuracy**.
- answer specific questions using transcriptome.

Materials and Methods

Learning dataset:
123 patients, tumour/non-tumour samples,
histo-pathological subtypes (MOBIT project, Niklinski 2017)

Machine learning:
Random-forest model (Breiman 2001)

Model testing:
Sensitivity/specificity and ROC curves

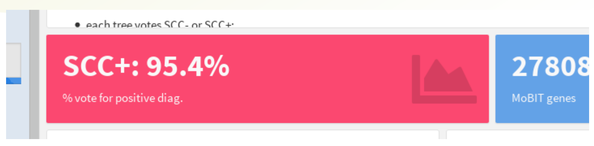
Render:
User friendly **web platform**
(R + shiny + flexdashboard)

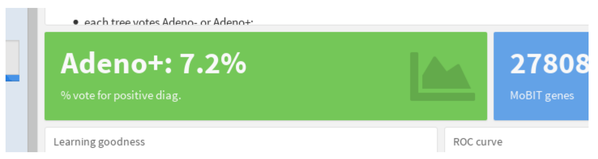
Results

Reactive modeling: **the model adapts** to the provided genes.

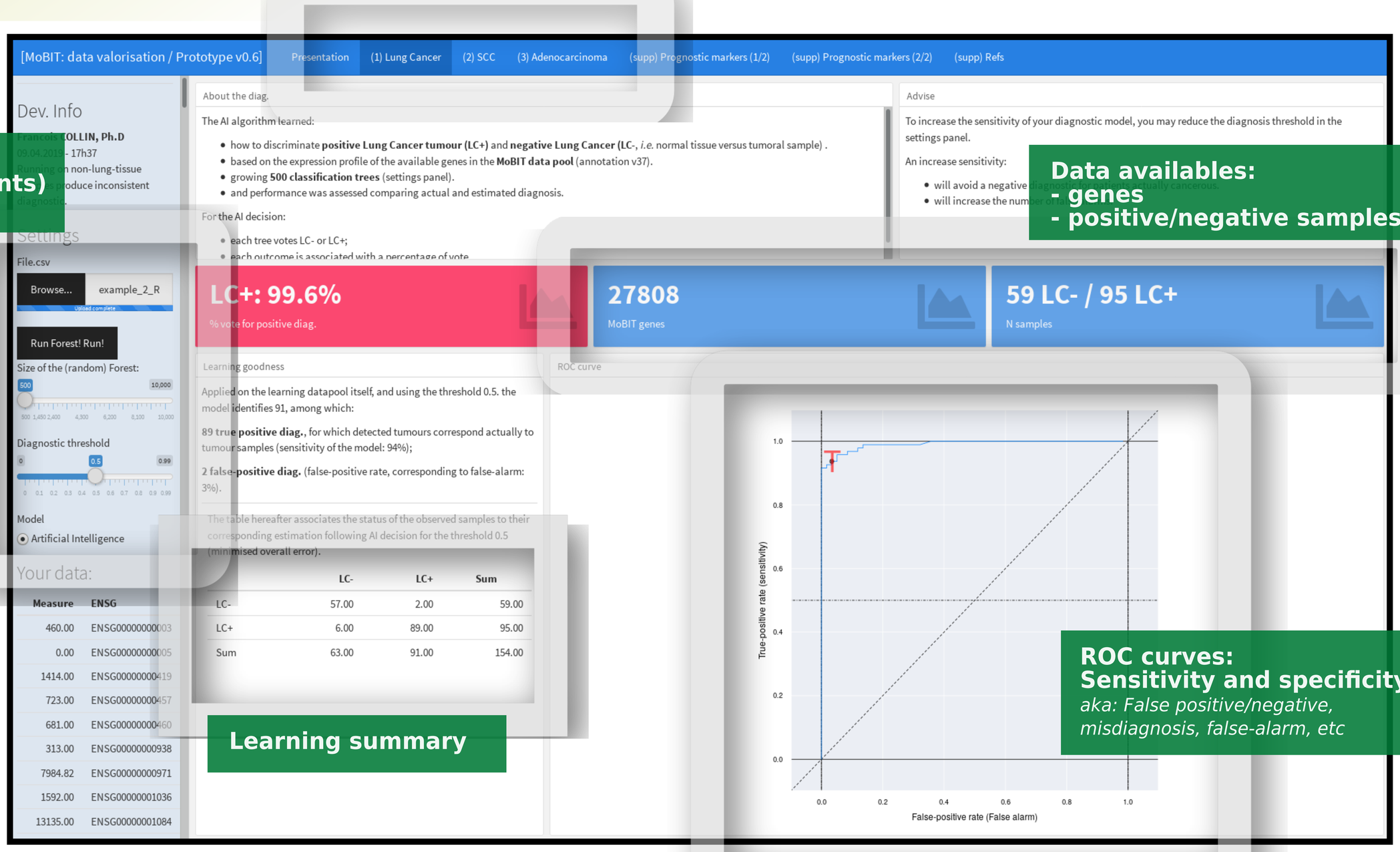
Best predictive results for whole transcriptome gene set.
Good performance for lower number of genes.

Predictions used for:
- diagnostic based on whole available gene set.
- hypothesis test (e.g. robustness of gene signature)
- hypothesis generation (e.g.identification of important genes)





3 binary diagnosis



Input data (genes/counts) & settings

Data availables:
- genes
- positive/negative samples

Learning summary

ROC curves:
Sensitivity and specificity
aka: False positive/negative, misdiagnosis, false-alarm, etc

Discussion

A simple use of complex data:
+ Phenotyping based on transcriptome analysis.
+ Easy comparison/external validation.

Perspectives for:
+ **personalised medicine,** subtypes based on gene expression
+ **prognostic** (depends on data availability on later stages of follow-up)

Machine learning and artificial intelligence can **increase usability of complex genomic data,** necessary step toward wider implementation of personalised medicine.

References

Niklinski J et al. "Systematic biobanking, novel imaging techniques, and advanced molecular analysis for precise tumor diagnosis and therapy: The Polish MOBIT project." Advances in medical sciences 62.2 (2017): 405-413.

Breiman, L. "Random forests." Machine learning 45.1 (2001): 5-32.

~~ No conflicts of interest to declare. ~~