

## Exploration of liver fibrosis based on machine learning analysis

*Nuria Perez-Alvarez<sup>1</sup>, Esteban Vegas<sup>2</sup>, Jose Àngel Jimènez Lasanta<sup>3</sup>, Raúl V Rodríguez Iniesta<sup>4</sup>, Cristina Pérez Balagoró<sup>5</sup>, Cristina Tural<sup>6</sup>*

<sup>1</sup>nuria.perez@upc.edu, Department of Statistics and Operational Research, Technical University of Catalonia-Barcelona Tech

<sup>2</sup>evegas@ub.edu, Department of Statistics, University of Barcelona

<sup>3</sup>jjimenez.germanstrias@gencat.cat, Radiology Department, Hospital Germans Trias i Pujol

<sup>4</sup>rodrigueziniesta@hotmail.com, Radiology Department, Hospital Germans Trias i Pujol

<sup>5</sup>cristinapbalaguero@gmail.com, Radiology Department, Hospital Germans Trias i Pujol

<sup>6</sup>ctural@flsida.org, Fight against AIDS Foundation, Hospital Germans Trias i Pujol

The assessment of a new diagnosis tool for the hepatic fibrosis is the main goal of this study. In order to achieve this objective, three different techniques were practiced to a cohort of patients: Acoustic radiation force impulse imaging (ARFI), the AST to Platelet Ratio Index (APRI) and the elastometry (ET), which is treated as the gold standard diagnosis technique. Furthermore, the data was explored to identify meaningful differences between patients in the liver diagnostic groups.

First of all, the concordance between techniques was assessed by means of scatterplots and the calculation of correlation coefficients. Multivariate analysis for the demographic and clinical variables was performed to identify outliers and to determine whether a subset of them can inform the profile of patients with liver fibrosis. Non supervised and supervised multivariate techniques, such as, principal component analysis (PCA), kernel PCA, random forests, CART and support vector machine techniques were applied.

The available data set contained demographic and clinical screening variables (a total of 105) belonging to 89 HIV infected patients. According of the approach features for some of the analysis only the continuous variables were used.

We were successful in identifying several differences between the metabolites found in normal liver and damaged liver by fibrosis, including cirrhosis. The results obtained by the different machine learning techniques are displayed and compared.

Machine learning approach may be applied to a large amount of clinical problems to describe patterns of patients presenting a specific disease.

**Keywords:** machine learning, multivariate analysis, biomarker, liver.