Simplifying the Interpretation of Steroid Metabolome Data by a Machine-Learning Approach

Tarik Kirkgoz[1], Semih Kilic[2], Zehra Yavas Abali[1], Ali Yaman[3], Sare Betul Kaygusuz[1], Mehmet Eltan[1], Serap Turan[1], Goncagul Haklar[3], Mahmut Samil Sagiroglu[4], Abdullah Bereket[1], Tulay Guran[1]

1)Marmara University, School of Medicine, Department of Pediatric Endocrinology and Diabetes, istanbul, Turkey
2)Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milan, Italy
3Marmara University, School of Medicine, Department of Biochemistry, istanbul, Turkey
4)Genpute Computation Technologies, istanbul, Turkey

P1-03

Background: Liquid chromatography-mass spectrometry (LC-MS) based panels of steroid hormones and their precursors offer a distinct pattern of steroid metabolome for various disorders of adrenal and gonadal steroidogenesis. However, it may not be easy to handle this high throughput data rapidly in clinical setting which requires expert opinion for correct interpretations. Analytical results of steroid panelling can be allied to automated review systems to simplify the complexity of data for disease-related interpretation.

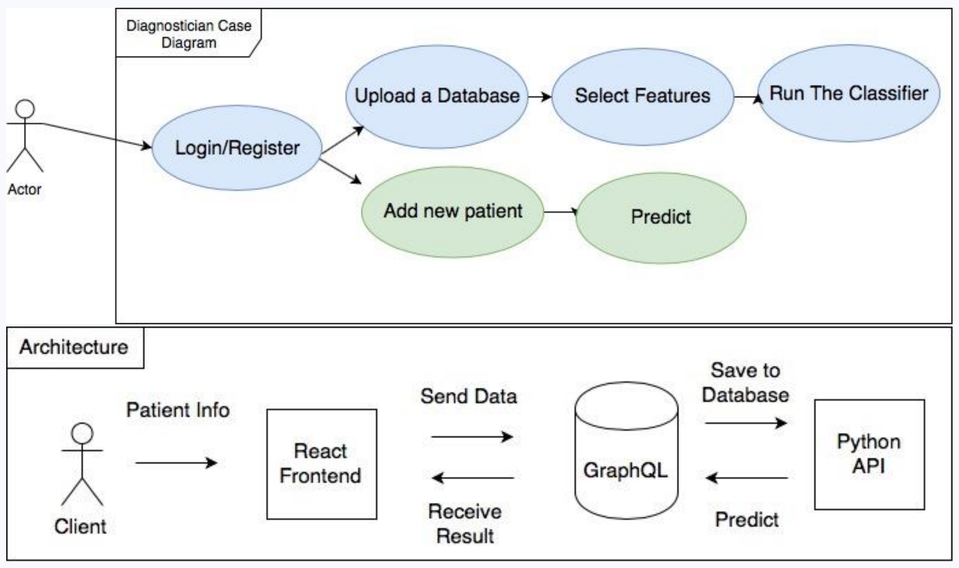
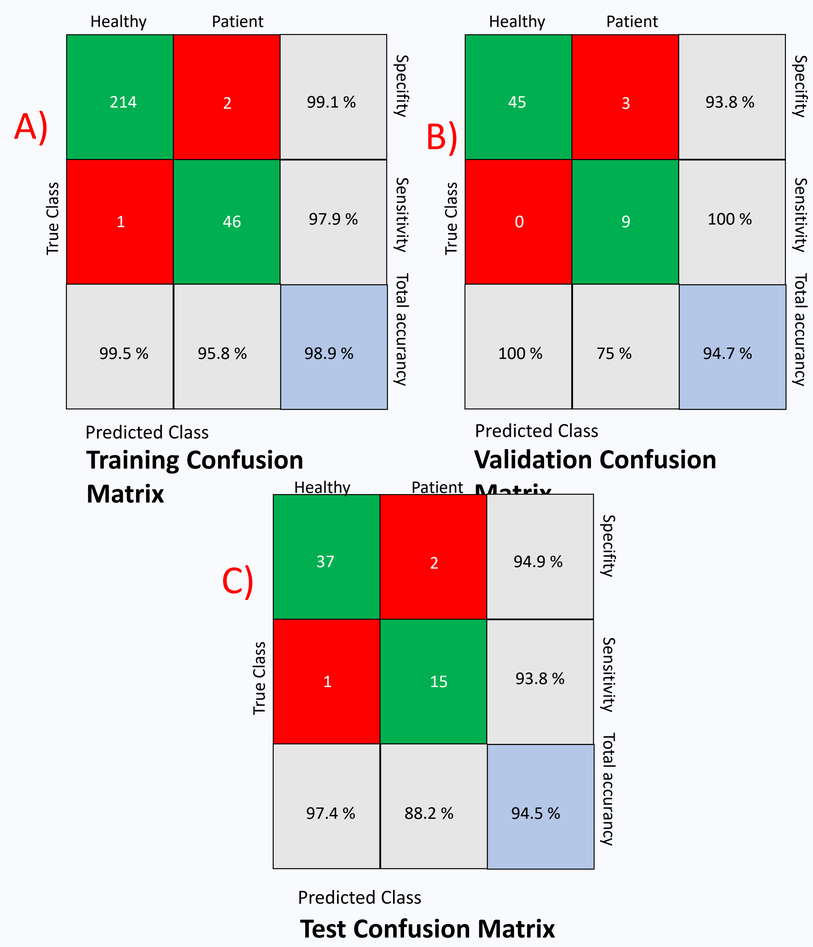


Figure 1) Schematization of programme

Methods: We have implemented a machine-learning algorithm for a time-saving and experience-independent review and interpretation of analytical results. We have tested the performance of this algorithm using our archived data of quantitation of 16 steroid hormones and precursors by an LC-MS/MS based panel in 500 healthy controls and 427 treatment-naive children with a disorder of adrenal steroidogenesis. This cohort included classic CYP21A2 (*n*=75), non-classic CYP21A2 (*n*=19), CYP11B1 (*n*=66), mutation-positive HSD3B2 (*n*=31), mutation-negative HSD3B2 (*n*=21), CYP11B2 (*n*=19), CYP17A1 (*n*=11), POR (*n*=7) deficiencies and non-CAH PAI (*n*=21). Due to the relatively low numbers of some of the conditions in the patient cohort, the number of samples in one class has outnumbered the other one. This imbalance has been over-comed by utilizing data sampling and boosting algorithms, specifically Random Oversampling Boosting (RUSBoost).



D)

True Class \ Predicted Class	CYP11B1	Non-Cah PAI	Non-Classic 21 OH'lase	POR	17 OH'lase	Aldosterone synthase
CYP11B1	43	3				1
Non-Cah PAI	1	19				
Non-Classic 21 OH'lase		1	17			
POR			1	5		
17 OH'lase					4	2
Aldosterone synthase				1	3	15

Figure 2) **A,B,C) Confusion matrixes and results**
D) Subgrouping patients data results

Results: Dataset of 415 patients fed to the algorithm with 10-fold cross validation to prevent overfitting. For discrimination of patients from the healthy controls; the sensitivity and specificity of the RUSBoost algorithm was 97.7% and 92.6%, respectively. The differentiation of each disorder could be achieved with overall accuracy of up to 95% independent of age and sex.

Conclusion :

- ✓ Application of RUSBoost machine learning algorithm enables a rapid and standardized review of complicated plasma steroid panelling data, which can widely be used by clinicians to make correct diagnosis for disorders of steroidogenesis.