

ASMS Conference.

June 2020

Machine Learning Models Predict Renal Cell Carcinoma Status from Multiplatform Urine-based Metabolomics.

Olatomiwa O. Bifarin^{1,2}, David A. Gaul⁵, Rebecca Arnold³, John A. Petros^{3,4}, Facundo M. Fernández⁵, Arthur S. Edison^{1,2}

¹Complex Carbohydrate Research Center, University of Georgia, Athens, GA, 30602 ²Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, 30602 ³Department of Urology, Emory University, Atlanta, GA 30342 ⁴Atlanta VA Medical Center, Atlanta, GA 30033 ⁵School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta GA 30332 (USA)





1. Background

- The best chance of surviving Renal Cell Carcinoma (RCC) is through early diagnosis, however the disease is characterized by asymptomatic progression. [1]
- Currently, RCC is identified through cross-sectional imaging and biopsies, the latter being highly invasive and riddled with sampling errors. [2][3]
- We aim to predict RCC status from urine metabolic profile using mass spectrometry and nuclear magnetic resonance.

4. Propensity Matching

Propensity matching and model cohort characteristics. Study cohort characteristics before matching and after matching (model cohort). p-values were calculated using the t-test. For unequal and equal sample sizes, Welch and Student t-test were used respectively.

Acknowledgements

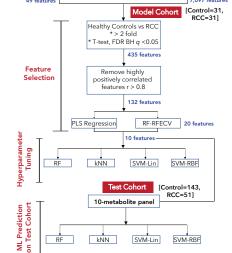




5. Machine Learning Pipeline

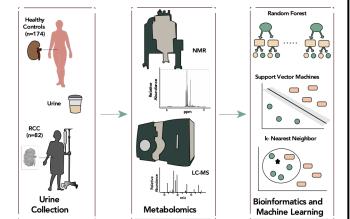
Normalized MS Data

Normalized NMR Data

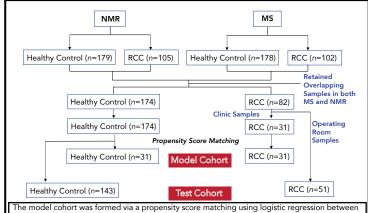


Using the model cohort, all NMR and MS features were subjected to a hybrid method of feature selection culminating in 10 selected metabolites. Hyperparameters for four different machine learning models were tuned using the model cohort and the 10-metabolite panel. Final predictions were made with using the test cohort under cross-validated conditions. PLS: Partial Least Squares. RF- RFECV: Random Forest Recursive Feature Elimination – Cross Validation FDR-BH: False Discovery Rate Benjamini Hochberg procedure k-NN: k-nearest neighbors. SVM: Support Vector Machines (Lin: linear, RBF: Radial Basis Function)

2. Graphical Abstract

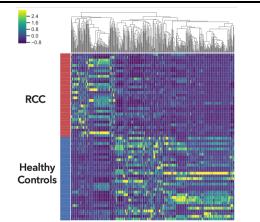


3. Patient Selection



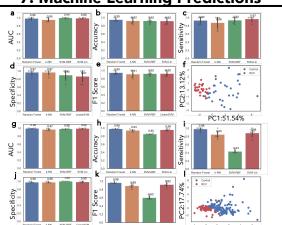
31 RCC samples and the best 31 matched healthy control samples. The second cohort, the test cohort composed of 51 RCC samples collected in the operating room and 143 healthy controls.

6. Differential Metabolites



Hierarchical clustering of 435 metabolic features with q values < 0.05 and > 2-fold change in the model cohort. z-scores are represented as shown in the color bar. Yellow represents higher level of metabolites while dark blue represents lower level of metabolites.

7. Machine Learning Predictions 8. F



Machine learning predictions using the 10 metabolites panel. (a-f) in the model cohort, and (g-l) in the test cohort. In the model cohort, random forest gave the best prediction with an accuracy of 95% under 5-fold cross validation conditions (b). In the test cohort, random forest gave the best prediction with an accuracy of 98% under 5-fold cross validation conditions (h).

8. Future Directions

Formula	KI [IIIIII]	Wiode	ı
C8 H9 N O	2.562	positive	720
NaN	6.290	positive	1481
C8 H19 N	3.449	positive	2102
C7 H18 N8 O6 S	1.133	positive	3141
NaN	1.184	positive	3675
C4 H12 N O6 P	2.595	positive	3804
NaN	4.049	positive	3872
C10 H21 N3 O8 P2 S	0.821	positive	4080
C9 H18 N9 O2 P	2.591	negative	6261
MS/MS experiments for identification of the 10			
metabolite panel			

9. References

- 1. Ann. Oncol. 2019, 30 (5), 706-720.
 - Radiol Clin North Am 2017, 55 (6), 1235-1250.
 - Curr Urol Rep 2017, 18 (4), 28.