Automated and interpretable machine learning for MS metabolomics: Predicting cancer diagnosis

LinkedIn

@BifarinPhD
@facundofGT

Georgia Institute of Technology

Olatomiwa O. Bifarin[1] and Facundo M. Fernández[1,2]

[1]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332, United States.
[2]Petit Institute of Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA 30332, United States.

Poster Tour

Fernández Lab

Background

The selection of optimal machine learning (ML) models for MS-based metabolomics is crucial but often involves tedious evaluation. Automated Machine Learning (AutoML) can automate this process, but the outputs can be difficult to understand, necessitating the need for complex model interpretation. AutoSklearn *[1]* was used for AutoML model selection, models were interpreted using the KernelSHAP method *[2]*, and the pipeline was tested on a renal cell carcinoma (RCC) urine-based metabolomics LC-MS dataset *[3]*.

Methods

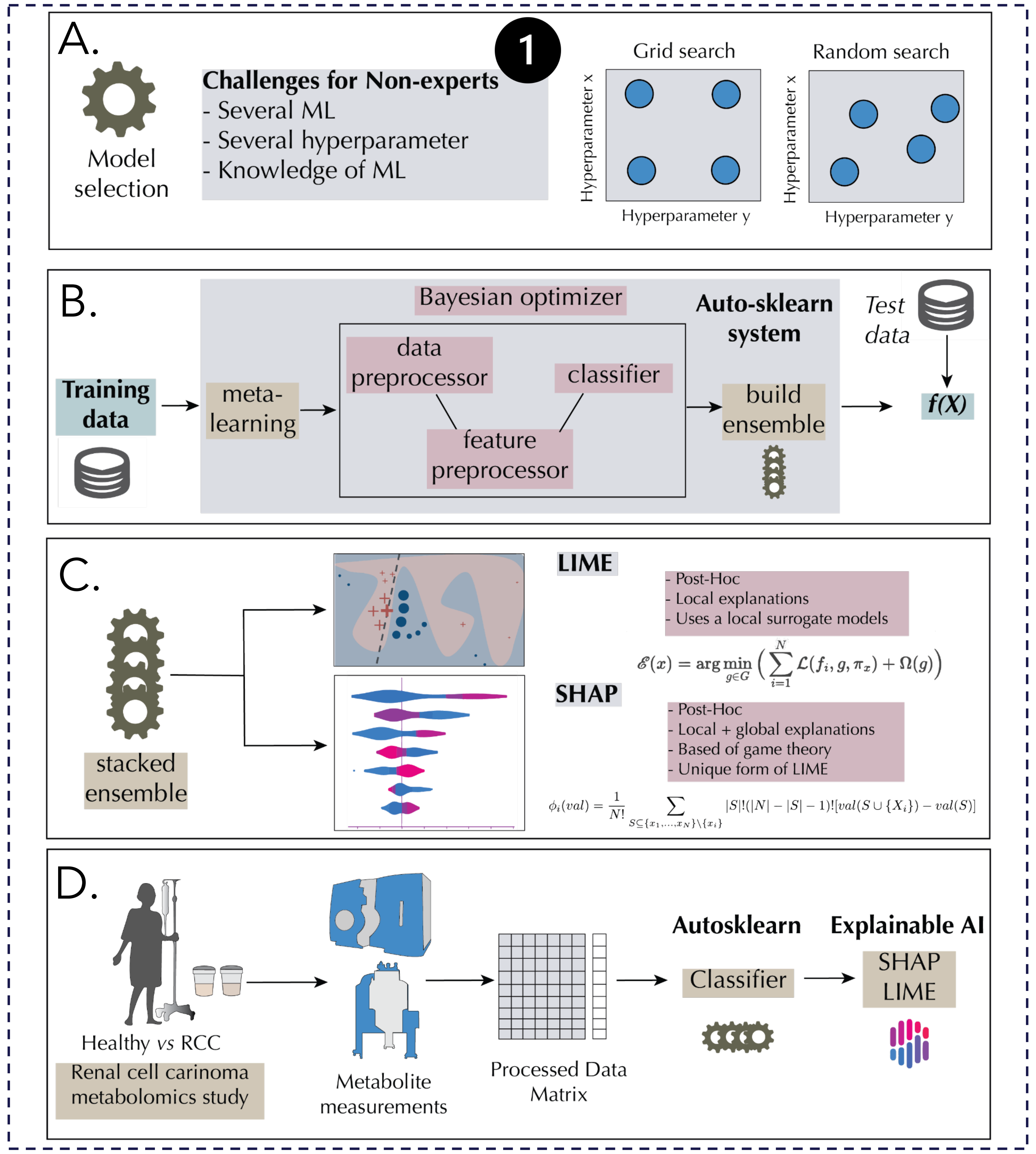

Fig 1: Automated & Interpretable Machine Learning Workflow. (A), Challenges for ML model selection. (B), Auto-Sklearn system. (C), Explainable AI (XAI) methods. (D), Applying AutoML and XAI to RCC study. Local Interpretable Model-agnostic Explanations – LIME, Shapley Additive exPlanation – SHAP

Methodology Details

$$\emptyset_j(val) = \frac{1}{N!}\sum_{S\subseteq\{1,\ldots,N\}\setminus\{j\}}|S|!\,(|N|-|S|-1)!\,[val(S\cup\{j\}) - val(S)] \quad (1)$$

$$\sum_{j\in N}\emptyset_j(val) = val(N) - val(\{-\}) \quad (2)$$

$$val(S\cup\{j\}) = val(S\cup\{k\}) \;\; \forall\, S\subseteq\{1,\ldots,N\}\setminus\{j,k\} \Longrightarrow \emptyset_j = \emptyset_k \quad (3)$$

$$val(S\cup\{z\}) - val(S) = 0 \;\forall\, S\subseteq\{1,\ldots,N\} \Longrightarrow \emptyset_z(val) = 0 \quad (4)$$

$$\mu_{\mathrm{sh}}(S) = \frac{N-1}{\binom{N}{|S|}|S|(N-|S|)} \quad (5)$$

$$\arg\min_{\emptyset_0,\ldots,\emptyset_n}\sum_{S\subseteq N}\mu_{\mathrm{sh}}(S)\left(\emptyset_0 + \sum_{i\in S}\emptyset_i - val(S)\right)^2 \quad (6)$$

Shapley value and kernel SHAP

Shapley value is a principled approach used to compute the individual contributions of elements within a cooperative system **(1)**, in this case, metabolomic features in a machine learning classification context. Shapley values guarantee fairness properties, namely: additivity **(2)**, symmetry/consistency **(3)**, and dummy **(4)**. Kernel SHAP is a combination of linear LIME + Shapley values. Using a weighted linear regression model as the local surrogate model and an appropriate weighting kernel **(5)**, the regression coefficients of the LIME surrogate model estimate the SHAP values. Finally, **(6)** shows how the Shapley values are estimated.

Results

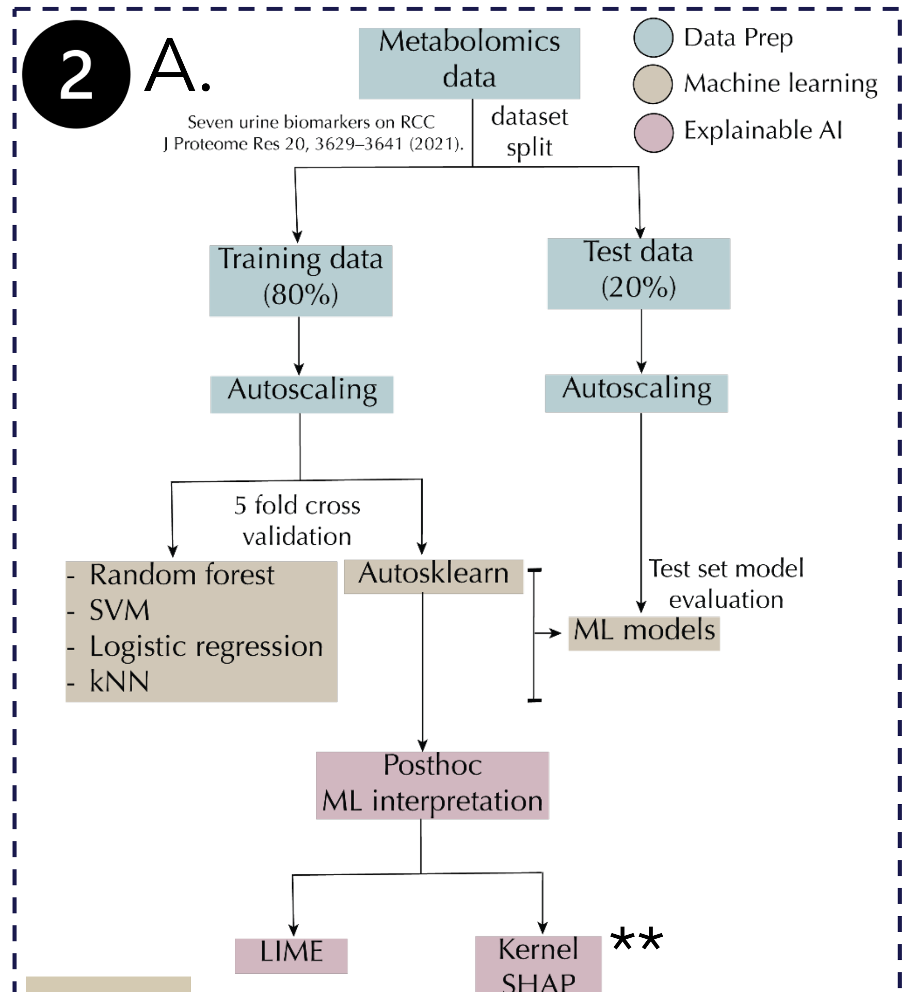

Table 1

Models	Training (ROC-AUC)	Test (ROC-AUC)
Random forest	0.98 ± 0.03	0.89
k-NN	0.96 ± 0.04	0.84
Logistic regression	0.97 ± 0.03	0.89
Support vector machine	0.97 ± 0.03	0.89
Autosklearn	0.97 ± 0.03	**0.97**

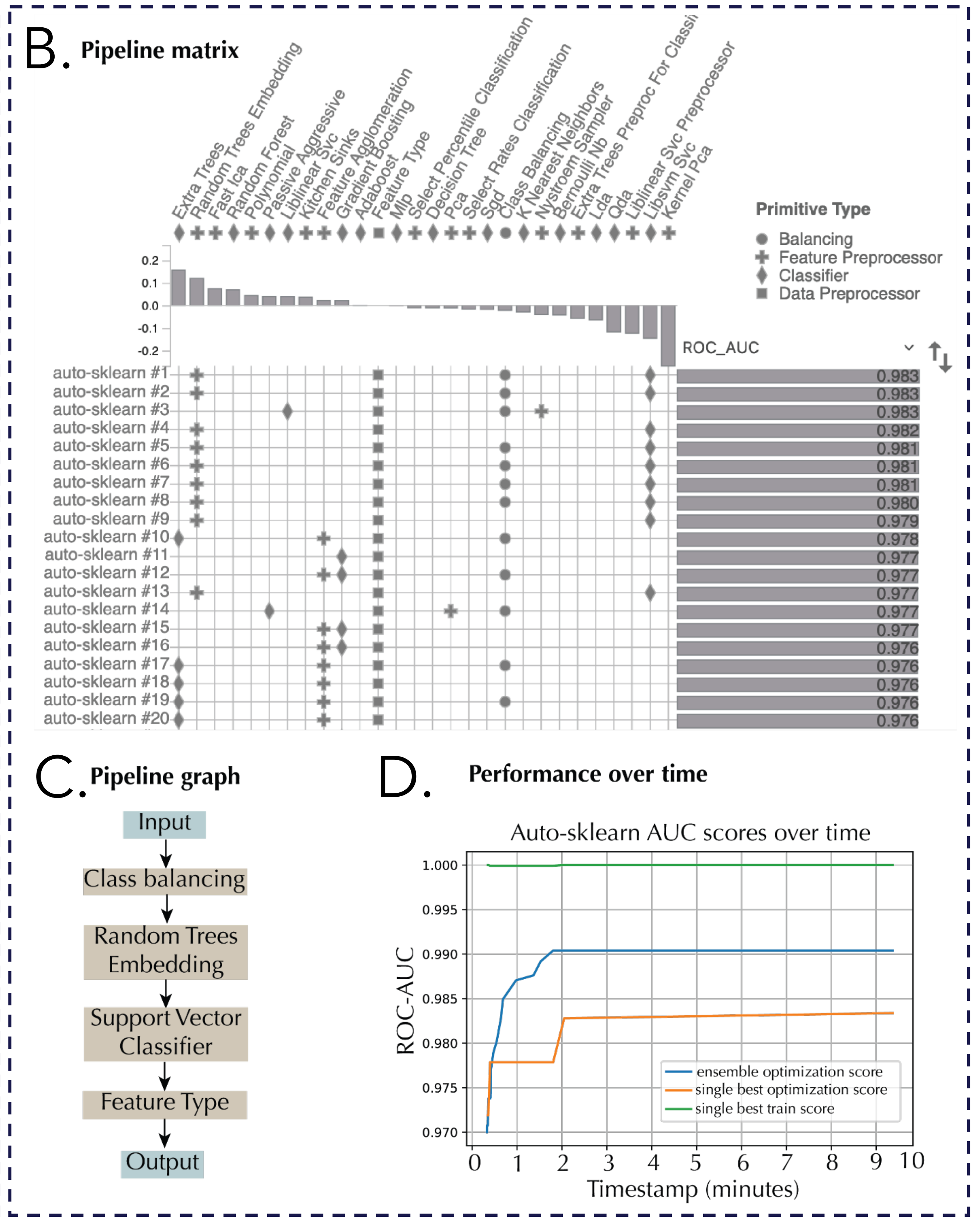

Fig 2: Computational Pipeline and Machine Learning Results. (A), Computational pipeline for data analysis. (B), Autosklearn pipeline matrix showing models and their primitives. (C), Autosklearn pipeline graph for the best model. (D), Performance over time, highlighting some metric scores during model training. Table 1: Machine learning results.

Results

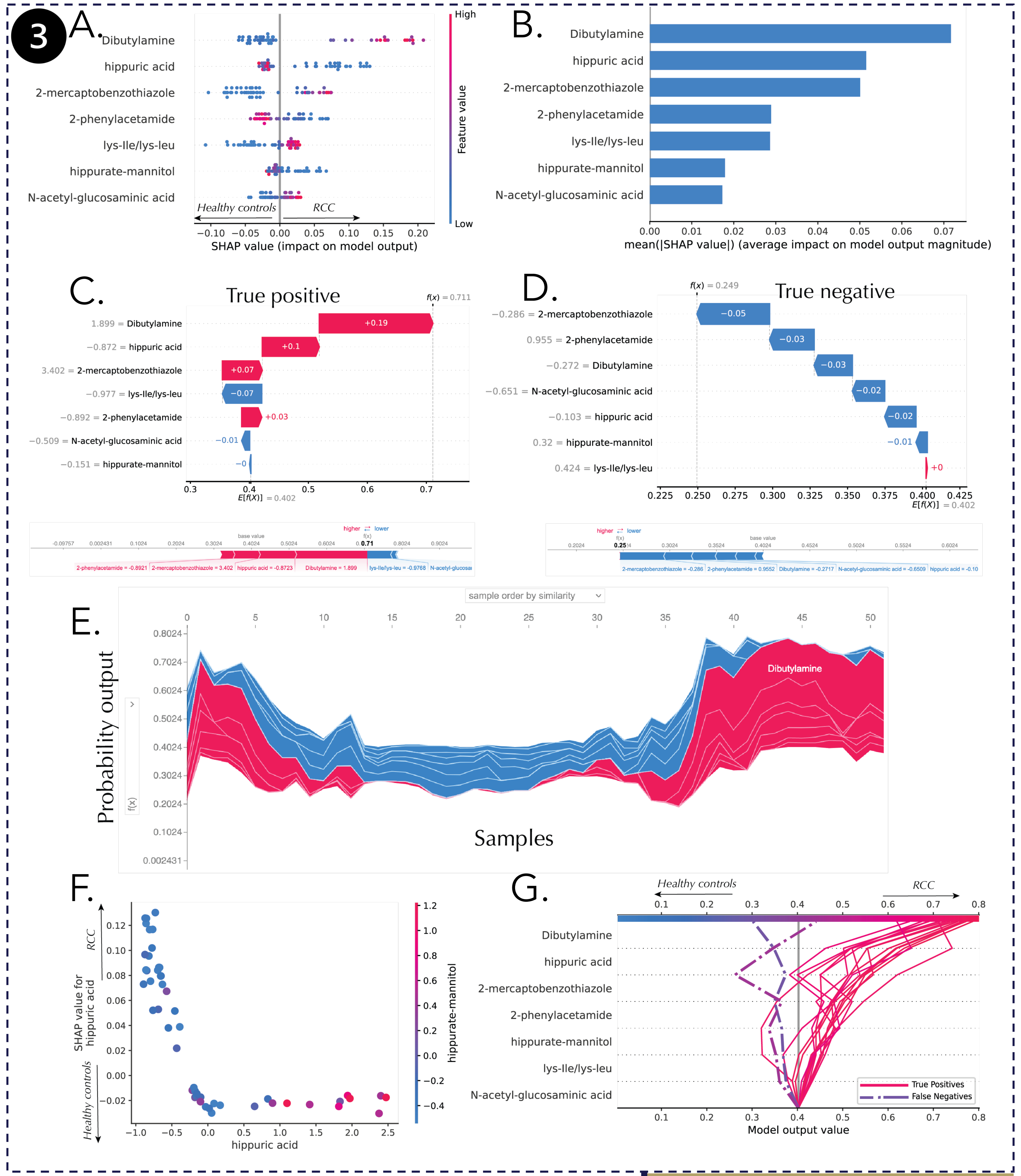

Fig 3: ML Interpretation of the AutoML Model used for the RCC Detection (Test set). (A), Beeswarm plot and (B), Summary plot showing global interpretation of the model. (C), Local explanation for true positive (RCC). (D), Local explanation for true negative (Healthy Controls). (E), Force plots for all test sets. (F), Dependence plot showing the interaction between hippuric acid and hippurate-mannitol derivative. (G), Decision plots highlighting true positives and false negatives.

References

[1]: Feurer et al., (arXiv, 2020) arXiv:2007.04074

[2]: Lundberg, S. & Lee, S.-I. Arxiv (2017) arXiv:1705.07874.

[3]: Bifarin, O. O. et al. J Proteome Res 20, 3629–3641 (2021).