XAItest: Evaluating XAI-Based Feature Discovery Methods

Ghislain Fievet, Julien Broséus, David Meyre, Sébastien Hergalant

XAItest: Evaluating XAI-Based Feature Discovery Methods.

Ghislain Fievet[1], Julien Broséus[1] , David Meyre[1] , Sébastien Hergalant[1]
[1]UMRS INSERM 1256 NGERE, 54500 Vandœuvre-lès-Nancy

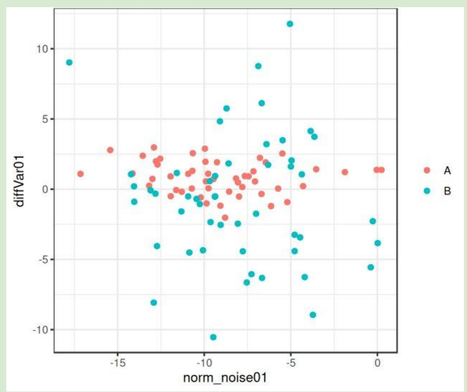
Identifying key variables in high-dimensional omics datasets remains a major challenge, particularly when relationships between features are non-linear, multimodal, or involve complex interactions. While classical statistical methods such as t-tests effectively capture mean differences, they often fail to detect more intricate patterns. Recent advances in explainable artificial intelligence (XAI) offer alternatives, but lacks systematic benchmarking of these methods.

In this study, we present XAItest, a framework designed to evaluate the ability of various machine learning algorithms and XAI methods to identify relevant features under diverse data configurations. We test six synthetic data scenarios, including variance shifts, bimodal distributions, XOR-like structures, concentric circles and non-linear regression patterns. Four ML models, Decision Trees, Random Forests, Support Vector Machines, and Multi-Layer Perceptrons, are combined with feature importance metrics such as Gini Decrease, Accuracy Increase, SHAP values, and Olden scores. To interpret feature importance in a statistically meaningful way, we apply p-value transformation methods including mProbes, PIMP, and our novel approach, simThresh, which estimates significance thresholds through iterative simulation.

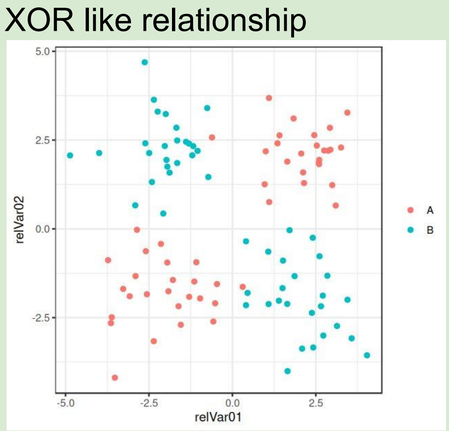
Our results show that all methods detect simple patterns reliably, while complex structures remain challenging for most algorithms. SimThresh emerges as a computationally efficient and effective method for setting statistical thresholds. This work provides a foundation for evaluation of XAI tools in biomarker discovery and is implemented as a Bioconductor package (10.18129/B9.bioc.XAItest).

6 scenarios undetected by classical statistical tests

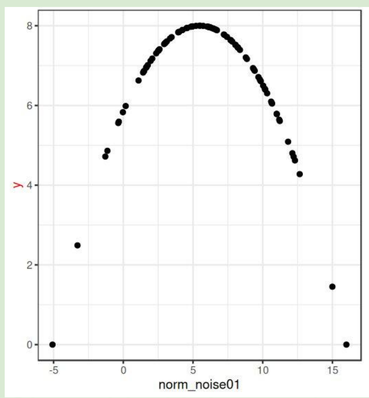
Variance differences



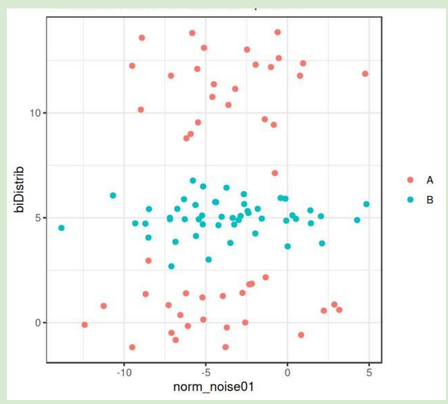
XOR like relationship



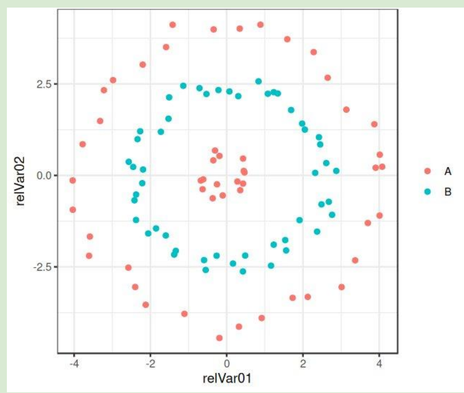
Parabole



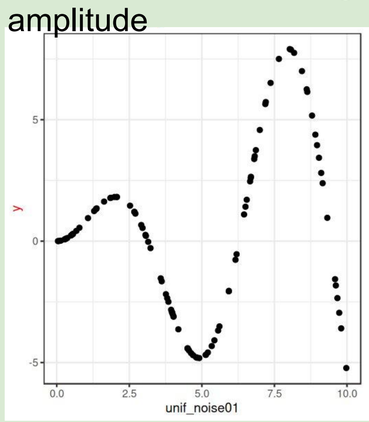
Bimodal distribution



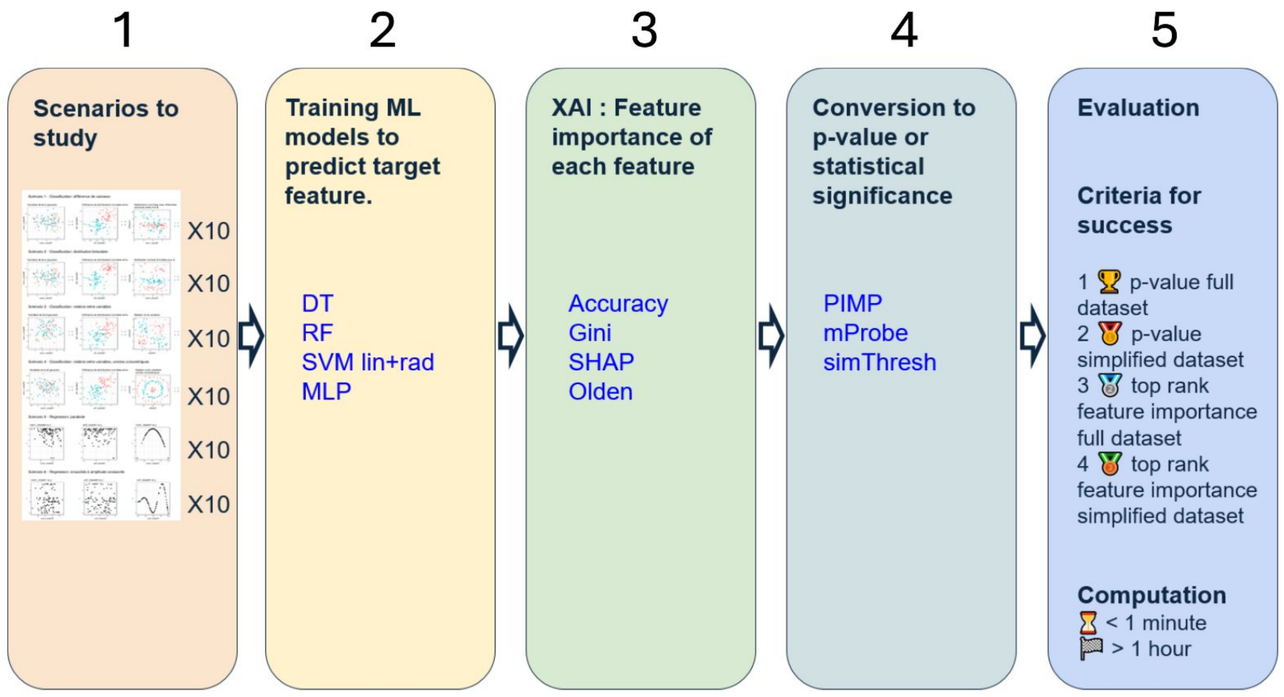
Concentric circles



Sinusoid with increasing amplitude



Workflow



PIMP, mProbes

Involves training multiple ML models to determine a null distribution

Returns a p-value for each feature

simThresh

Inserts a variable calibrated to match a specified target p-value.

Involves training a single ML model.

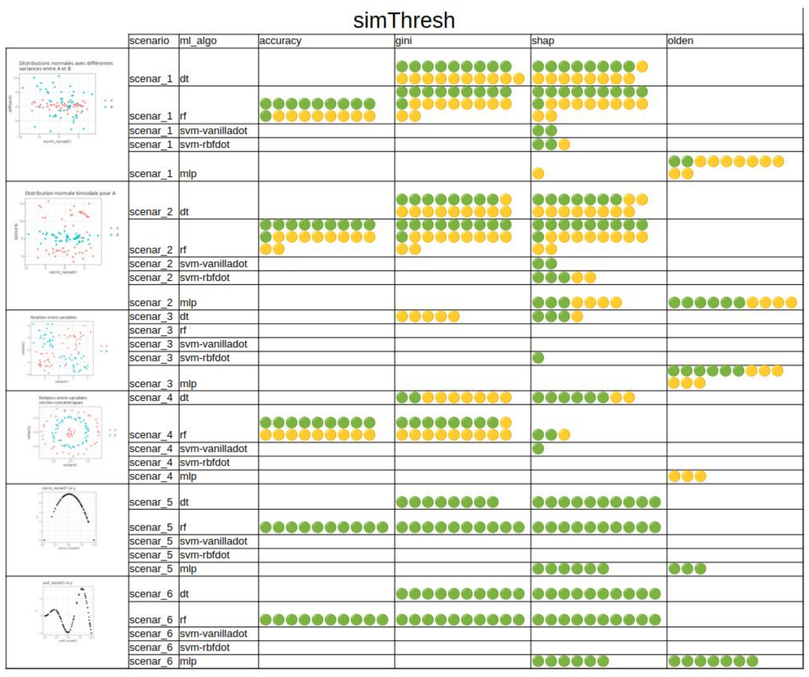
Returns, for each feature, whether the significance threshold is reached or not.

Recommended combination

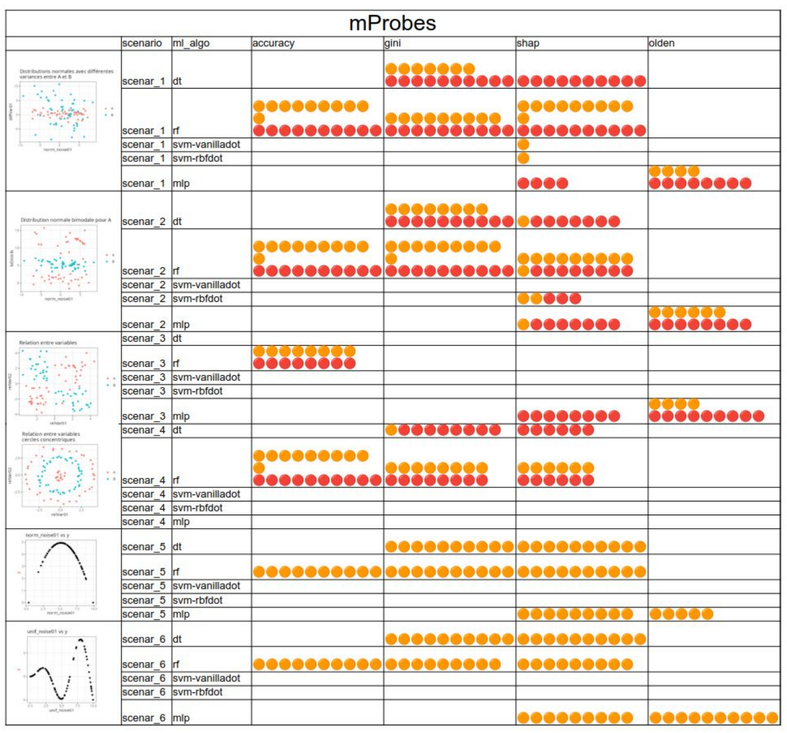
DT + Gini + simThresh

- the markers identified by p-value threshold exactly match the scenario markers using the full dataset.
- the markers identified by p-value threshold exactly match the scenario markers using the reduced dataset.
- the top-ranked markers by p-value exactly match the scenario markers using the full dataset.
- the top-ranked markers by p-value exactly match the scenario markers using the reduced dataset.

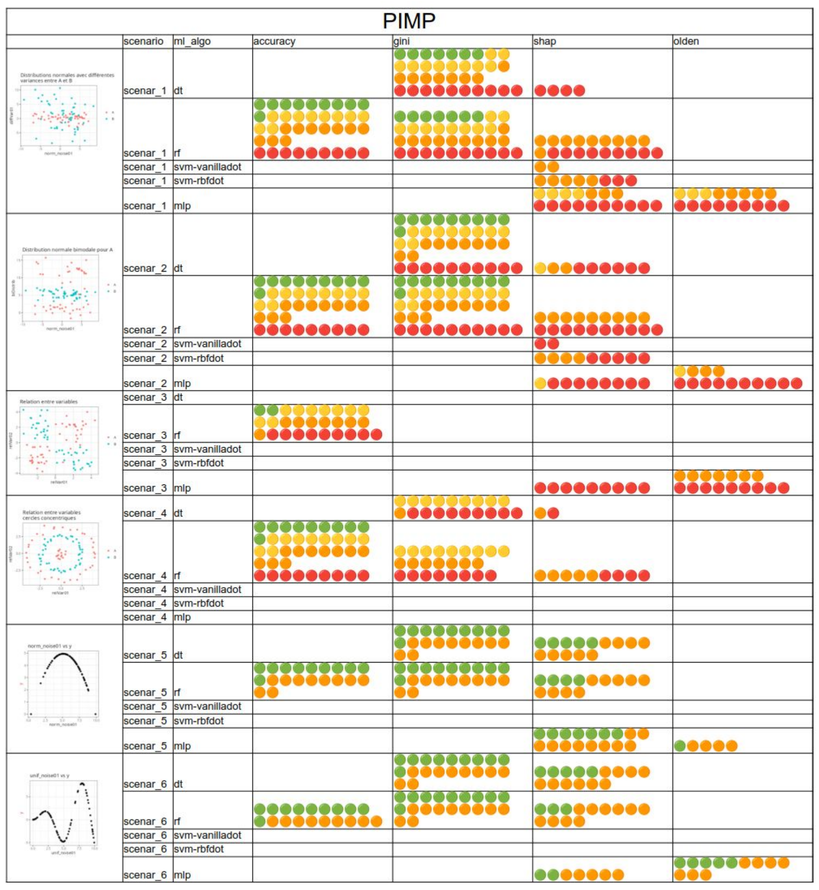
simThresh



mProbes



PIMP



References

[1] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, “Permutation importance: a corrected feature importance measure,” Bioinformatics, vol. 26, no. 10, pp. 1340–1347, May 2010, doi: 10.1093/bioinformatics/btq134.

[2] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts, “Statistical interpretation of machine learning-based feature importance scores for biomarker discovery,” Bioinformatics, vol. 28, no. 13, pp. 1766–1774, Jul. 2012, doi: 10.1093/bioinformatics/bts238.

Contact ghislain.fievet@univ-lorraine.fr , ghislain.fievet@gmail.com

XAItest https://www.bioconductor.org/packages/devel/bioc/html/XAItest.html