ENHANCING SHAP EXPLANATION INTERPRETABILITY USING SUBGROUP DISCOVERY

MAËLLE MORANGES, THOMAS GUYET

AIstroSight, HCL, UCBL, Inria

INTRODUCTION

Predictive AI in Medicine: High-performing predictive models are often opaque. Their clinical adoption requires reliable and comprehensible explanations.

Objective: producing explanations that are:
- Clinically interpretable
- Model-agnostic
- Integrating interactions between variables
- Accounting for individual variability
- Both global and local

Proposed approach: Generation of explicit IF-THEN rules by combining SHAP and Subgroup Discovery [1].

Challenges:
- Explanations faithful to the internal behavior of the model
- Precise explanations beyond simple importance scores
- Alignment of XAI explanations with clinical reasoning

RELATED WORK

SHAP[2]: Reference XAI method in medicine

Advantages:
- Local and global explanations
- Attractive visualizations

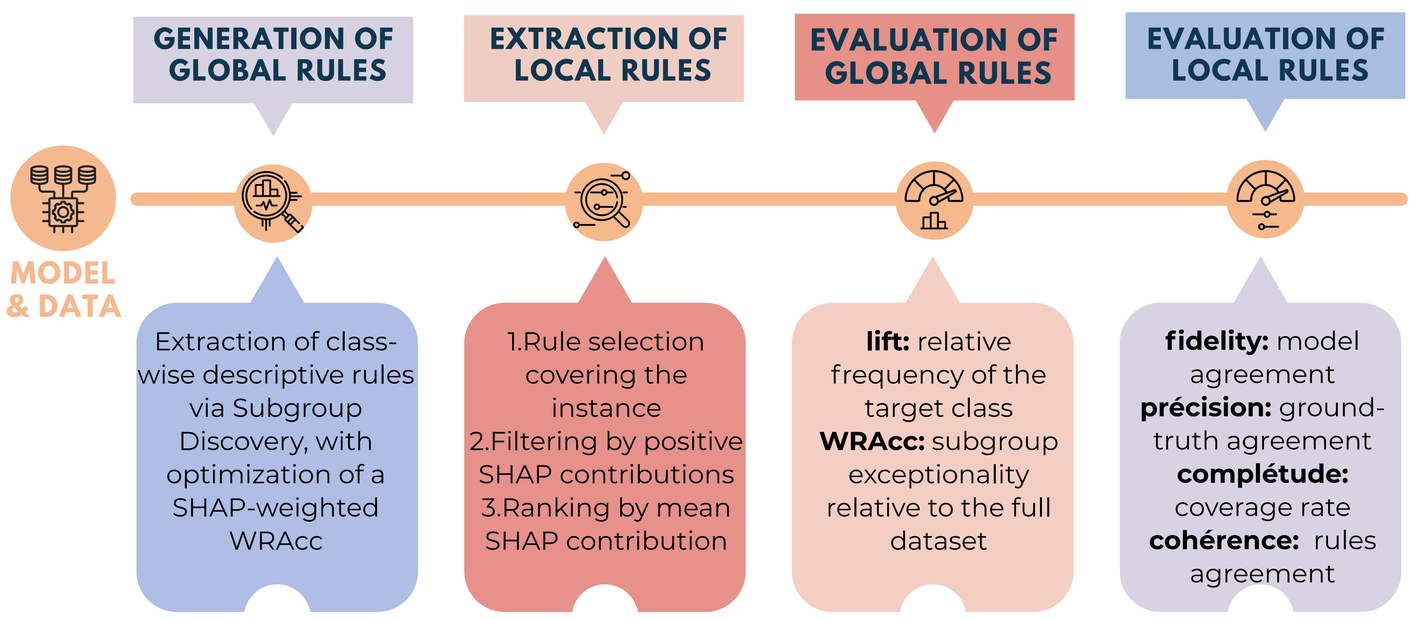
Limitations:
- SHAP values remain abstract for clinicians
- Lack of variable interactions
- Global averages mask patient diversity

Existing rule-based approaches[3,4,5]:

Advantages:
- Close to clinical reasoning
- Description of specific patient profiles
- Capturing variable interactions

Limitations:
- Only local or global explanations
- Rules are often long and therefore complex

METHOD



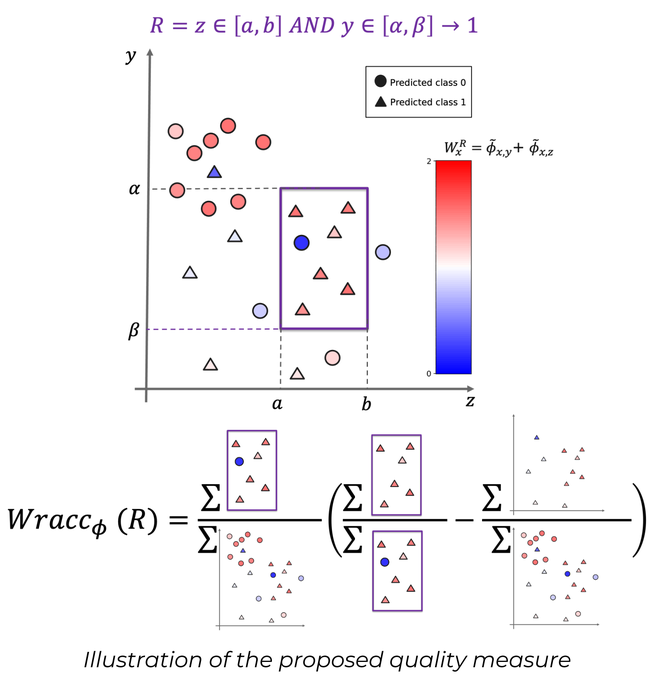
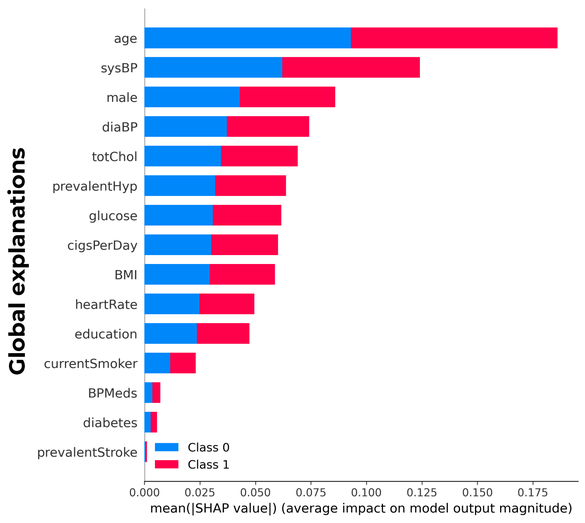


Illustration of the proposed quality measure

EXEMPLE OF RESULTS

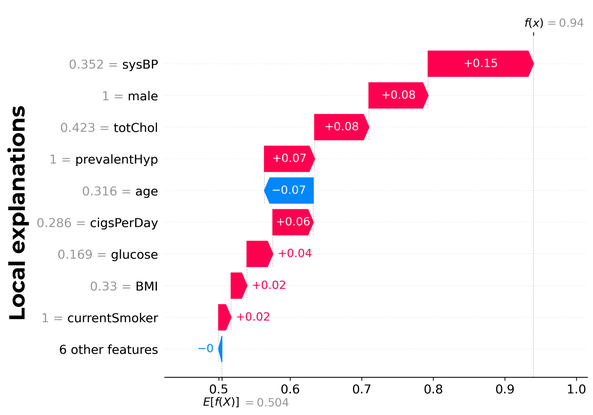
Comparison of global (top) and local (bottom) explanations on the Framingham dataset, contrasting standard SHAP (left) with our method (right)



Standard global SHAP visualization

$\text{WRAcc}_\phi(R)$	R	c
0.084	prevalentHyp=0	$\rightarrow 0$
0.076	male=0 AND prevalentHyp=0	$\rightarrow 0$
0.076	male=0	$\rightarrow 0$
0.073	age < 0.26	$\rightarrow 0$
0.071	diabetes=0 AND prevalentHyp=0	$\rightarrow 0$
0.071	BPMeds=0 AND prevalentHyp=0	$\rightarrow 0$
0.069	prevalentHyp=0 AND prevalentStroke=0	$\rightarrow 0$
0.063	BPMeds=0 AND male=0	$\rightarrow 0$
0.062	diabetes=0 AND male=0	$\rightarrow 0$
0.061	age $\in [0.26 : 0.42[$	$\rightarrow 0$
0.084	prevalentHyp=1	$\rightarrow 1$
0.082	age ≥ 0.74	$\rightarrow 1$
0.077	sysBP ≥ 0.32	$\rightarrow 1$
0.076	male=1	$\rightarrow 1$
0.069	prevalentHyp=1 AND prevalentStroke=0	$\rightarrow 1$
0.068	age $\geq 0.74 AND prevalentStroke = 0$	$\rightarrow 1$
0.066	age $\geq 0.74 AND diabetes = 0$	$\rightarrow 1$
0.065	prevalentHyp=1 AND sysBP ≥ 0.32	$\rightarrow 1$
0.064	prevalentStroke=0 AND sysBP ≥ 0.32	$\rightarrow 1$
0.062	diabetes=0 AND prevalentHyp=1	$\rightarrow 1$

Global rules extracted by our method (10 per class)



SHAP local explanation for the first test instance

$\phi_x(R)$	R	c
0.2174	prevalentHyp=1 AND sysBP ≥ 0.32	$\rightarrow 1$
0.1473	sysBP ≥ 0.32	$\rightarrow 1$
0.0828	male=1	$\rightarrow 1$
0.0701	prevalentHyp=1	$\rightarrow 1$

Local explanation generated by our method for the same instance

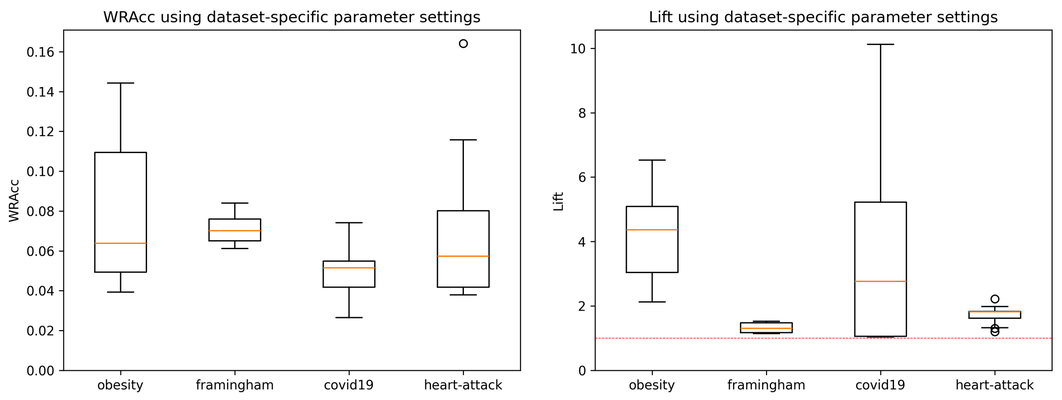
EXPERIMENTS

Dataset	|Classes|	|Features|	|Instances|	Model	Model Accuracy
Framingham	2	15	3,658	Random Forest	0.9758
Heart-attack	2	8	2,111	Decision Tree	0.9924
Covid19	2	19	1,048,575	Logic Regression	0.9384
Obesity	7	15	2,111	MultiLayer Perceptron	0.8511

Summary of the datasets used for evaluation, including the number of classes, features, and instances, as well as the predictive models and their corresponding accuracies

Dataset	depth	result_set_size	Fidelity	Accuracy	Completeness	Consistency
Framingham	2	10	0.9	0.88	0.97	0.8
Heart-attack	2	10	0.96	0.95	1	0.98
Covid19	2	20	0.92	0.88	0.92	0.99
Obesity	3	10	0.76	0.69	0.87	0.68

Optimal parameters (maximum rule depth and number of rules per class) and corresponding evaluation metrics (fidelity, accuracy, completeness, and consistency) for each dataset



Boxplots of rule-level quality measures obtained with the selected parameters for each dataset

REFERENCES

[1] *Wrobel. 1997. An algorithm for multi-relational discovery of subgroups. In ECML PKDD. Springer.*
[2] *Lundberg et Lee. 2017. A unified approach to interpreting model predictions.*
[3] *Ribeiro et al. 2018. Anchors : High-precision model-agnostic explanations. AAAI.*
[4] *Guidotti et al. 2019. Factual and counterfactual explanations for black box decision making. IEEE IS.*
[5] *Yuan et al. 2022. Visual exploration of ML model behavior with hierarchical surrogate rule sets. IEEE TVCG.*

CONTACT

Email: maelle.moranges@inria.fr
Website: https://mmaelle.github.io/
Date: 28/01/2026